SSE-DP-2023-3

統計的学習

(講義スライド)

- Statistical Learning-

国友直人・趙宇・湯浅良太（訳者）

Trevor Hastie and Robert Tibshirani (原著者)

統計数理研究所

2023 年 5 月

SSE-DP-2023-3

Lecture Slides of Statistical Learning

- by Trevor Hastie and Robert Tibshirani -

Translated by

Naoto Kunitomo, Yu Zhao and Ryota Yuasa

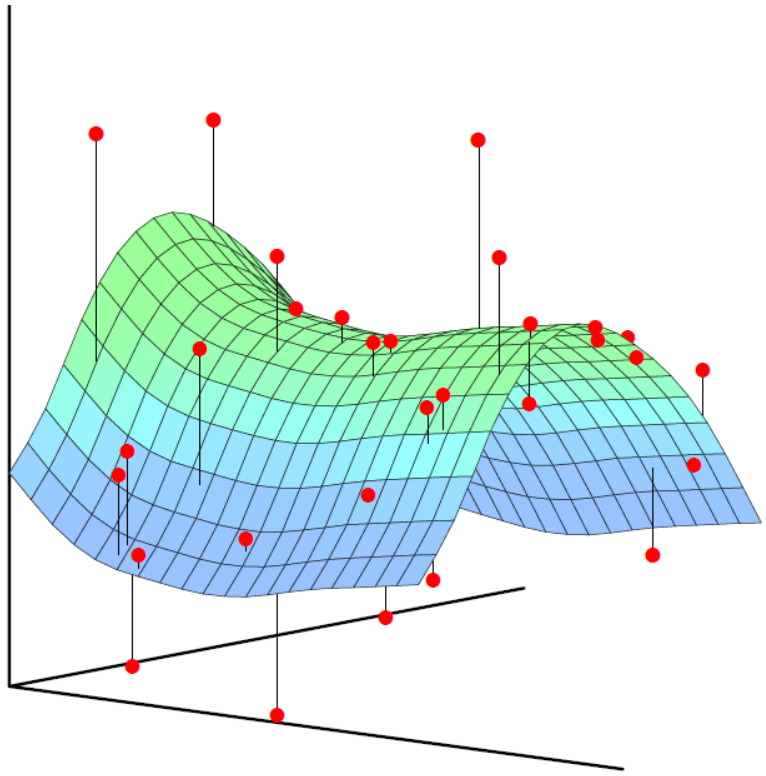
The Institute of Statistical Mathematics

May 2023

(Summary)

The lecture slides by Professors Trevor Hastie and Robert Tibshirani at Department of Statistics, Stanford University have been translated into Japanese with their generous permission. They would facilitate the teaching or delivering lectures on Statistical Learning in Japan, which is the growing field in Statistical Sciences.

統計的学習・講義スライド
-Statistical Learning-



（日本語版）国友直人・趙宇・湯浅良太
2023年5月版
（原著者）Trevor Hastie and Robert Tibshirani

日本語版・制作者からの序文

この(日本語)スライド講義録は元々は米国スタンフォード(Stanford)大学統計学科のヘイスティ(Hastie)教授とティブシラニ(Tibshirani)教授が同大学学部・大学院修士課程における講義の為に準備した英文スライドを(Hastie教授のご厚意により次頁のような許可を受け)日本語に翻訳したものである。なおこの日本語版では原スライドの誤植を修正、また幾つかの箇所で授業を行う上で有益と思われる補足を加えた。(翻訳の担当は国友1,2,3,7,11,日本版注;趙4,5,8,9;湯浅6,10,12,13の各章とし,その後に内容を調整した。)

統計数理研究所では「統計エキスパート人材育成」の為に大学統計教員育成センターを新たに立ち上げ、日本の大学学部専門課程・大学院修士課程における統計学教育を充実するための教材を開発中であり、この翻訳もそうした教材開発の一環として行われたもので、公開する。大学・大学院における統計教育の一助になれば幸いである。

2023年5月
国友直人（日本語版・作成者代表, 統計数理研究所）

(Trevor Hastie先生からの連絡)

An e-mail from Professor Trevor Hastie

April 5, 2023

Dear Professors Kunitomo and Takemura :

I asked my coauthor Prof. Rob Tibshirani, and he and I are both in agreement that we are fine for you to do what you propose. In particular, in (iii), we are happy for you to provide for free the Japanese versions of the slides and figures. In appropriate place you would put something like “Japanese conversion and translation done with permission of Trevor Hastie and Robert Tibshirani"

Best wishes
Trevor

目次

第1章 ニュース番組で取り上げられる統計学
- Statistics in the news -

- IBMのワトソン研究所：IMBスーパー・コンピュータ
- ニューヨークタイムズ August 5, 2009
- 選挙動向の予想：シグナルとノイズ(訳者注：翻訳あり)

統計的学習の幾つかの事例

- 前立腺癌のリスクファクター
- 対数ピリオドグラムから記録された音素
- 人口変数・食事変数・臨床検査にもとづき心臓発作を予測
- スパムeメールの検知
- 手書きの郵便番号の識別
- 遺伝子表現型にもとづいて組織サンプルから癌
- 給料と人口変数との関係
- 衛星画像のピクセルから利用用途

教師あり学習と教師なし学習

第1章 ニュース番組で取り上げられる統計学
- Statistics in the news -

IBMのワトソン研究所：IMBスーパー・コンピュータはすごい！(*Jeopardy*-playing supercomputer)

by Dawn Kawamoto DailyFinance 02/08/2011

誤りからの学習

Learning from its mis-takes

According to David Ferrucci (PI of Watson DeepQA technology for IBM Research),

ワトソン研究所のソフトウエアは自然言語処理の扱いに適する.

“機械学習(*machine learning)*に基づき課題への答えを導いたり, 解答が正しいか間違えかを学ぶことにより, 計算機はより賢く, スマートになる.”

For Today's Graduate, Just One Word: Statistics (人類学と考古学の大学・卒業生に必要なスキル：統計学)

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

Enlarge This Image

Thor Swift for The New York Times

Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

Multimedia

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

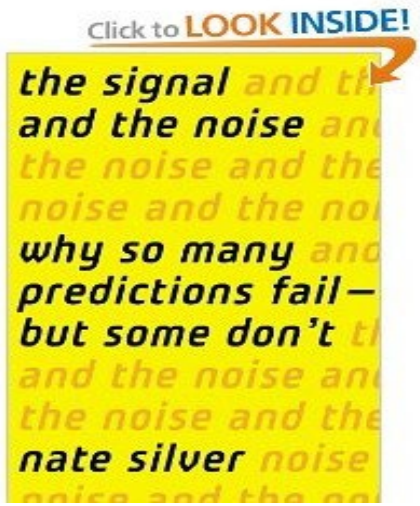
Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

ある日の新聞からの引用，
ニューヨークタイムズ
AUGUST 5, 2009
"I keep saying that the sexy job (素敵な仕事) in the next 10 years will be statisticians(統計家). And I'm not kidding (冗談を言っている訳ではない)."
— ハル・バリアン
HAL VARIAN,
グーグル主任エコノミスト
chief economist at Google.

記事：選挙動向の予想：シグナルとノイズ(訳者注：翻訳あり)

多くの選挙予想は外れた，しかし幾つかは上手く予想できた．ネイ・シルバーによるオバマの再選 2012年



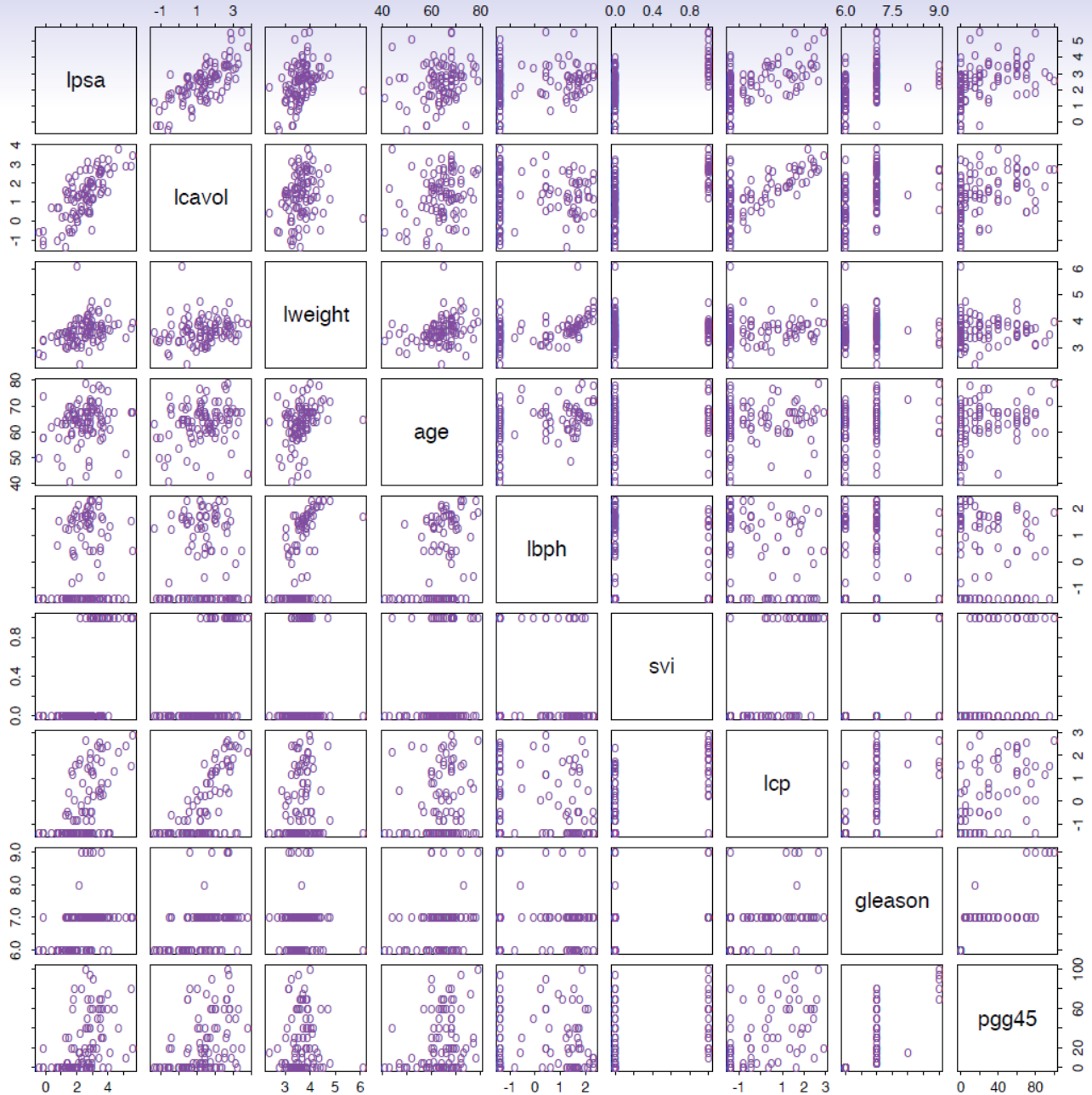
(訳者注：2016年D. Trump氏の事例？)

統計的学習の幾つかの事例

- 前立腺癌のリスクファクターを識別する(Identify the risk factors for prostate cancer).
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

lpsa:前立腺特異抗原水準の対数
lcavol:腫瘍径の対数
lweight:前立腺体重の対数
lbph:良性前立腺量の対数
svi:転移量
lcp:浸透率の対数
gleason:スコア値

(訳注)医学専門用語については訳者など非専門家には正確な理解は困難である。



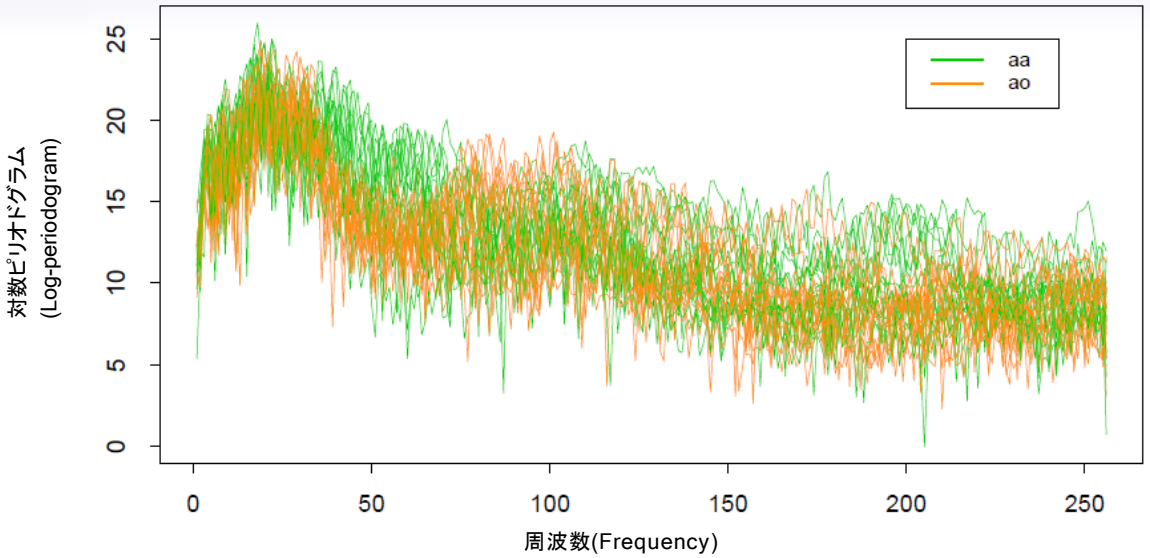
lpsa:前立腺特異抗原水準の対数
lcavol:腫瘍径の対数
lweight:前立腺体重の対数
lbph:良性前立腺量の対数
svi:転移量
lcp:浸透率の対数
gleason:スコア値
pgg45:gleasonスコアが4,5以上の割合

(訳注)医学専門用語については訳者など非専門家には正確な理解は困難である。

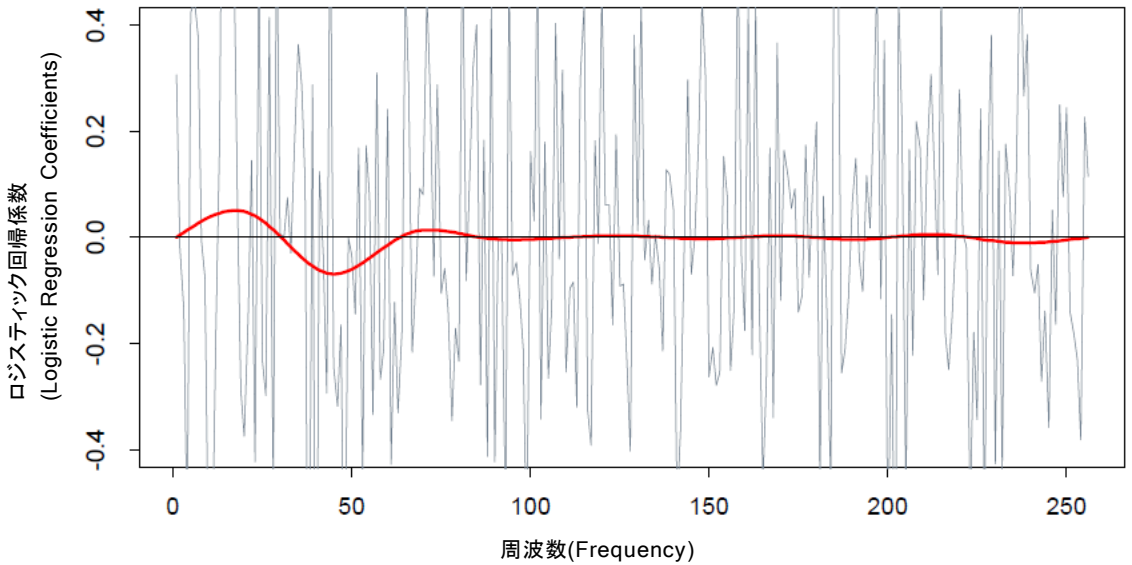
統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- 対数ピリオドグラムから記録された音素を分類する(Classify a recorded phoneme based on a log-periodogram).
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

音素（Phoneme）の例



音素の分類：原・制約付きロジッティック回帰

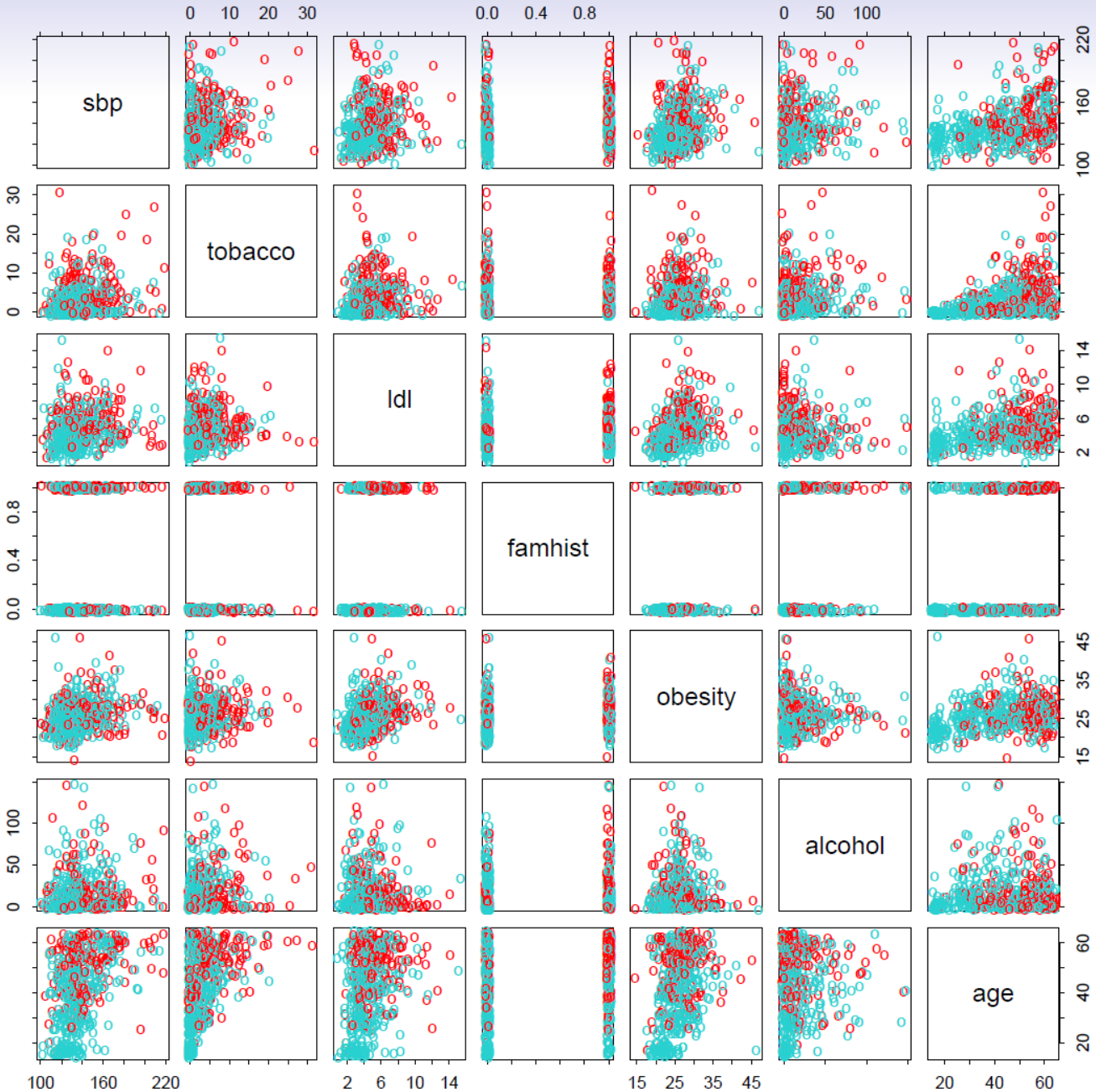


統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- 人口変数・食事変数・臨床検査にもとづき心臓発作を予測する

(Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements).

- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.



sbp:収縮期血圧
ldl:コレステロール
obesity:肥満
age:年齢

(訳注)医学専門用語については訳者など非専門家には正確な理解は困難である。

統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- スパムeメールの検知システムの制作

(Customize an email spam detection system).

- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.

スパムの検知(Spam Detection)

- ある個人への4601のメールデータ (名前George, HPラボ, 2000以前). 各データにラベル *spam* あるいは *email*.
- 目標: カスタマイズしたスパム・フィルターの設計.
- 入力の特徴量: メール中の 57の相対頻度(よく使われる単語(words)と句読点(punctuation marks).

	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

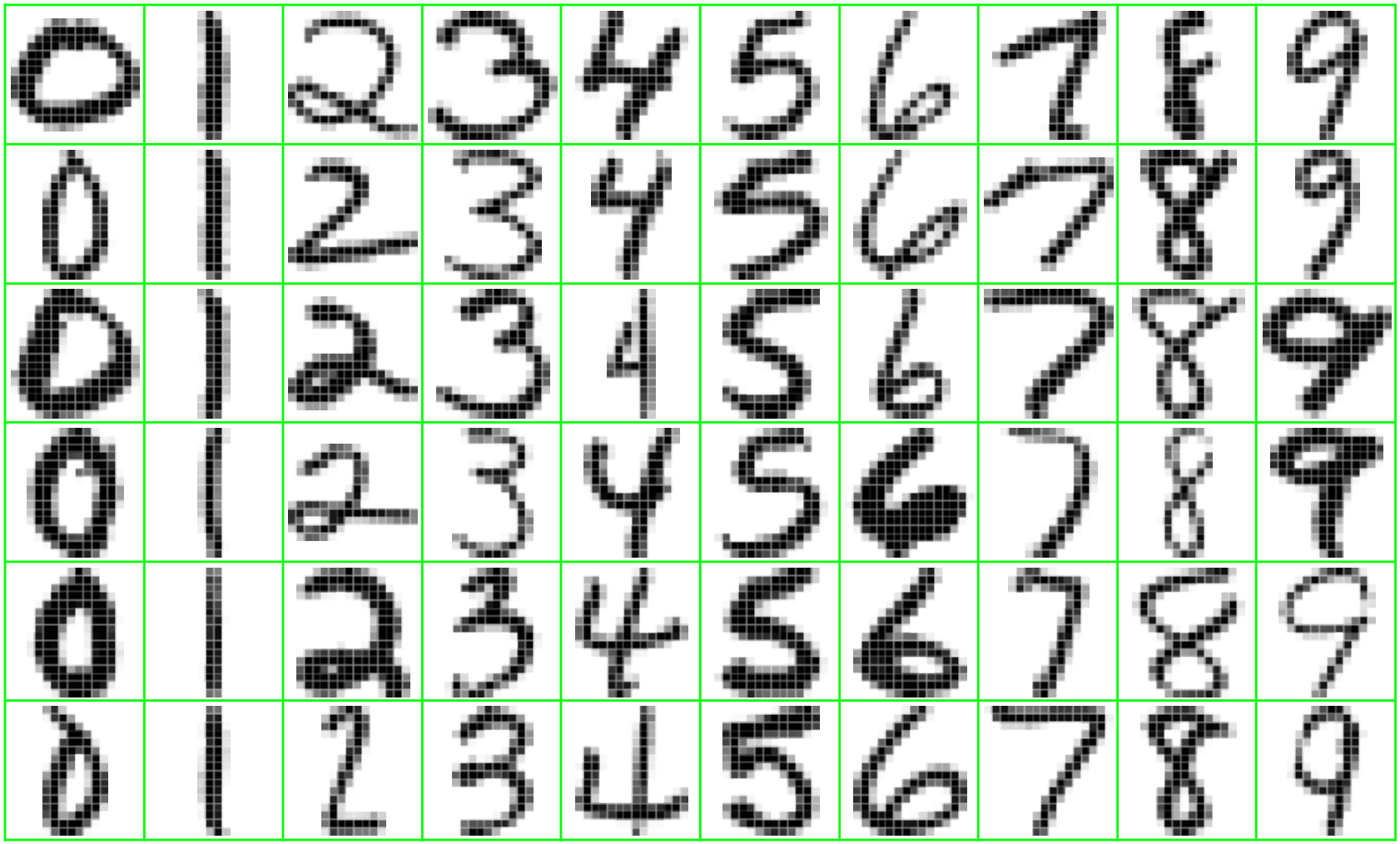
Emailメッセージ中の単語や文字の平均値. 単語や文字はspamとemail間で最大の差が見られるように選んでいる.

統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- 手書きの郵便番号の識別

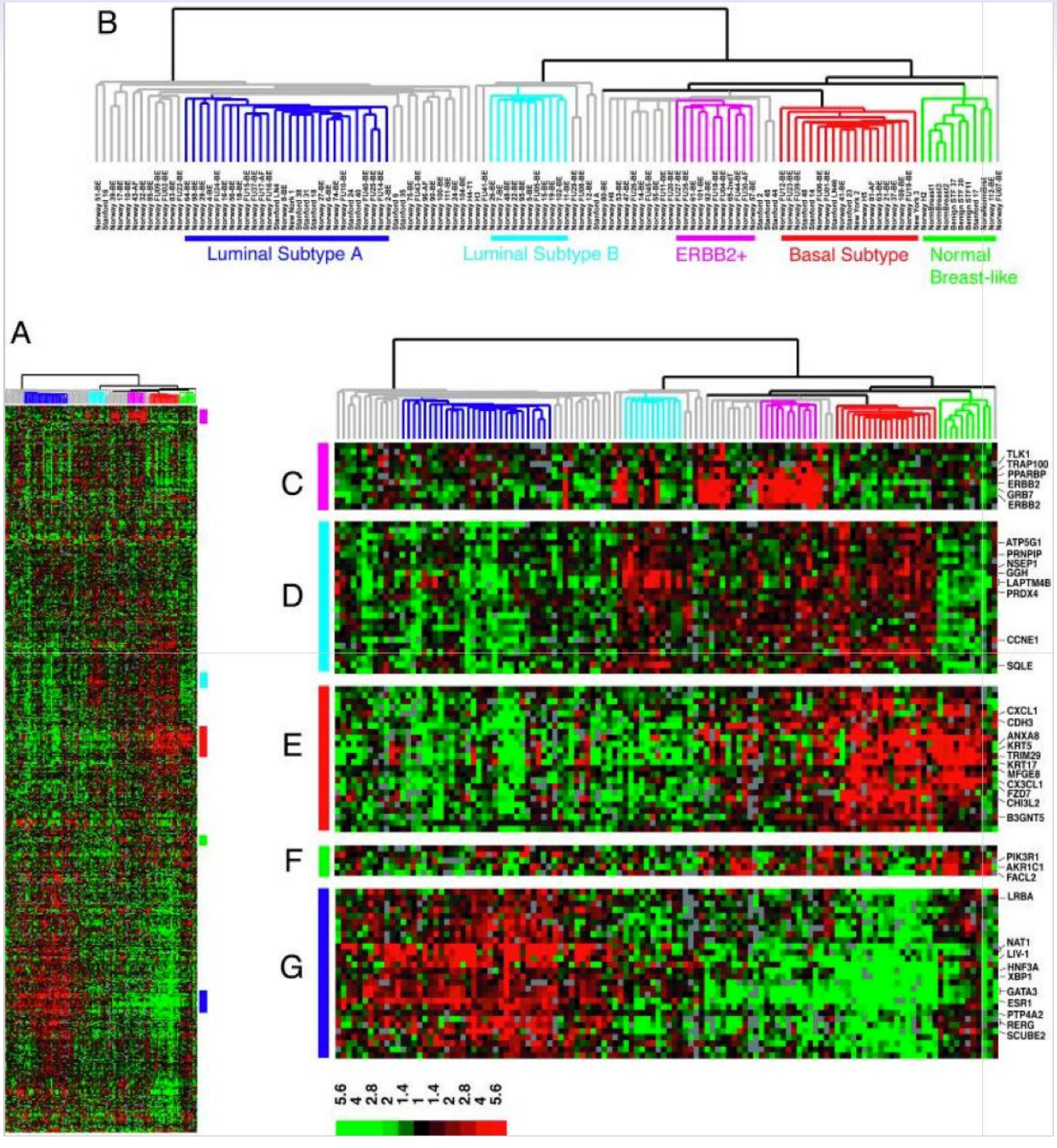
(Identify the numbers in a handwritten zip code).

- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.



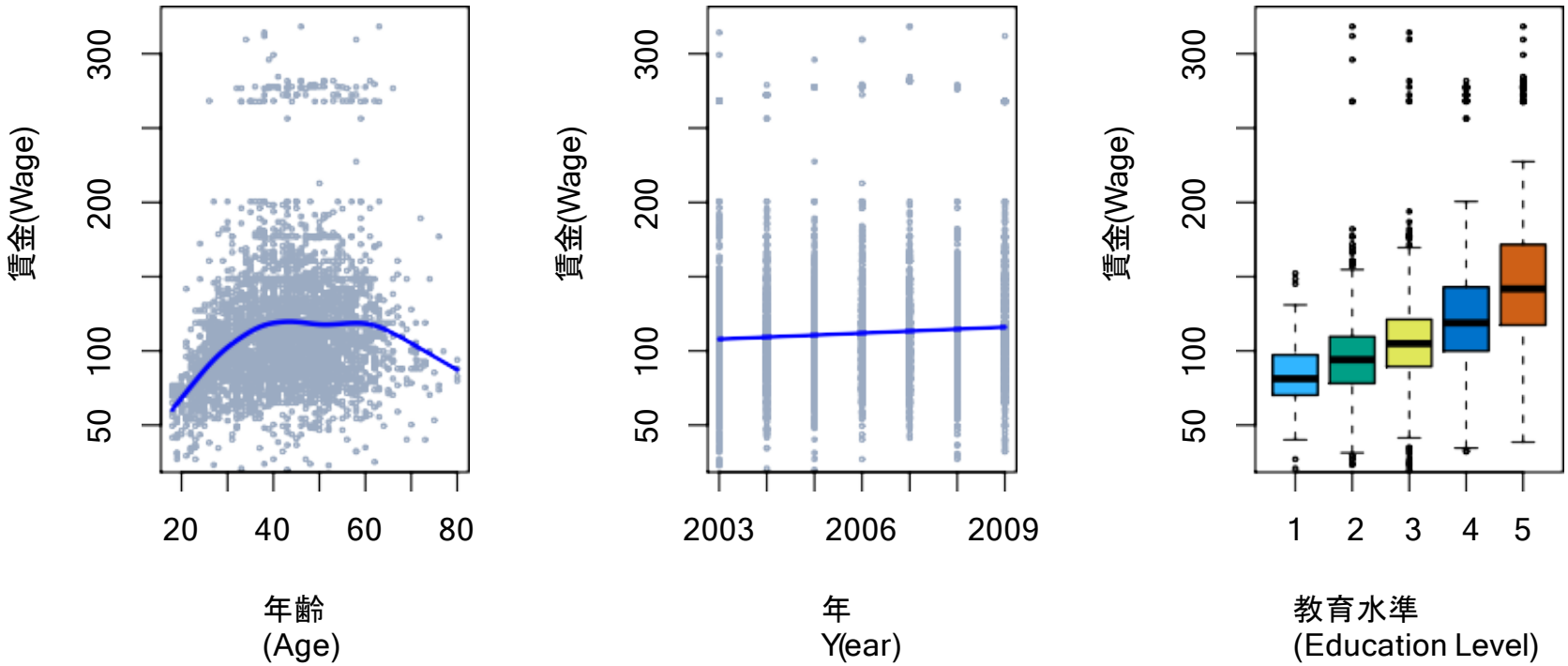
統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- 遺伝子表現型にもとづいて組織サンプルから癌クラスに分類する(Classify a tissue sample into one of several cancer classes, based on a gene expression profile).
- Establish the relationship between salary and demographic variables in population survey data.
- Classify the pixels in a LANDSAT image, by usage.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GGH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CHI3L2
B3GNT5
PIK3R1
AKR1C1
FACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6

統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- 人口調査データに基づく給料と人口変数との関係の分析 (Establish the relationship between salary and demographic variables in population survey data).
- Classify the pixels in a LANDSAT image, by usage.

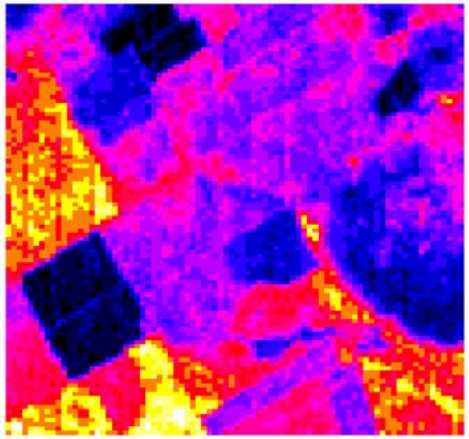


2009年米国の中央大西洋地域における男性に対する所得調査データ(Income survey data for males from the central Atlantic region of the USA in 2009).

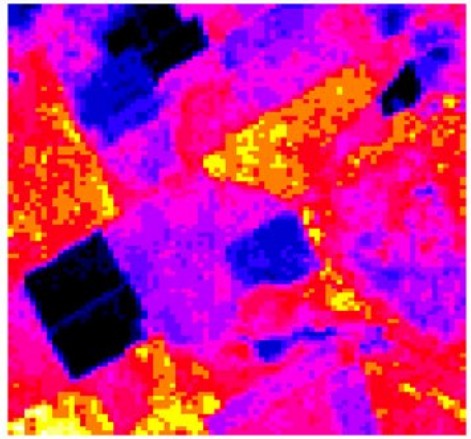
統計的学習の幾つかの事例

- Identify the risk factors for prostate cancer.
- Classify a recorded phoneme based on a log-periodogram.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profile.
- Establish the relationship between salary and demographic variables in population survey data.
- 衛星画像のピクセルから利用用途による分類(Classify the pixels in a LANDSAT image, by usage).

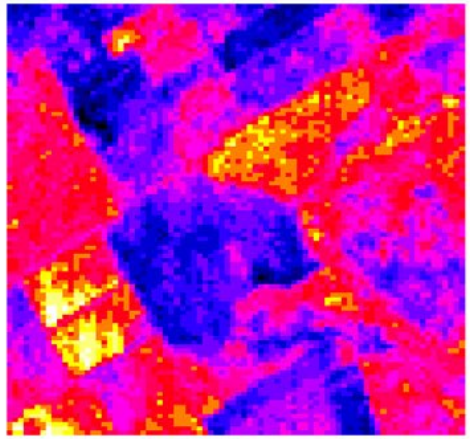
スペクトル領域1
(Spectral Band 1)



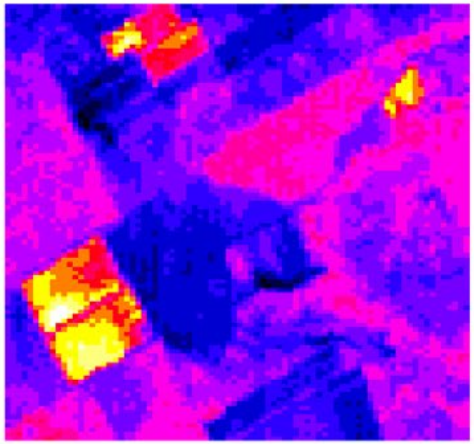
スペクトル領域2
(Spectral Band 2)



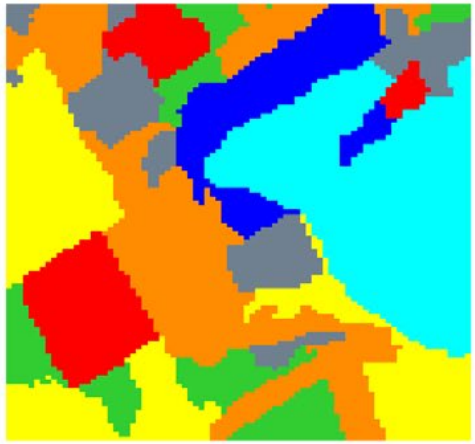
スペクトル領域3
(Spectral Band 3)



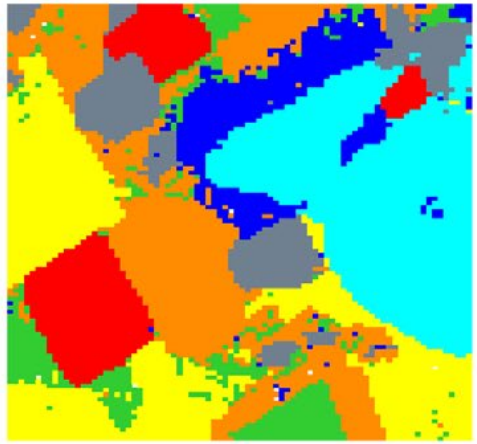
スペクトル領域4
(Spectral Band 4)



利用形態
(Land Usage)



予測した利用形態
(Predicted Land Usage)



利用形態(Usage) ∈{赤土red soil, ワタ畑cotton, 切り株畑vegetation stubble, 混合土mixture, 灰色土gray soil, 湿った灰色土damp gray soil}

教師あり学習問題
-Supervised Learning Problem-

出発点(Starting point):

- 目的変数(Outcome measurement) Y (しばしば 従属変数(dependent variable),反応変数(response),標的変数(target)などと呼ばれる).
- p個の予測変数ベクトル X (あるいは インプット(inputs), 回帰変数(regressors), 共変量(covariates), 特徴量(features), 独立変数(independent variables) などとも呼ばれる).
- 回帰問題(*regression problem)* ではY *は量的変数(*quantitative) (例えば価格, 血圧 など).
- 分類問題(*classification problem)*ではYは有限数の値のどれか, 順序付けられない集合 (生存/死亡, 0-9, ガン細胞標本のクラス など).
- トレーニングデータ $(x_1, y_1), \ldots, (x_N, y_N)$. 計測された観測値(標本, 事例など)とする。

分析目的

訓練(training)データから次のことを行いたい:

- まだ得られていないテスト事例を正確に予測したい

(Accurately predict unseen test cases).

- どの入力が出力・結果に影響したか、どのように影響したか知りたい

(Understand which inputs affect the outcome, and how).

- 予測や推測の質を評価したい

(Assess the quality of our predictions and inferences).

統計的考え方

- 様々な統計的方法をいつ、どのように利用するかはその背後にあるアイデアを理解することが重要.
- まずはより簡単な方法を理解することから、より複雑な方法を理解することにしよう.
- ある方法がどのくらい上手く、あるいは上手く行かないかパフォーマンスを正確に評価することが重要 [単純な方法はしばしば派手な(fancier)方法と同じくらい良いことがある!
- 統計的学習はエキサイティングな研究分野であり、科学・産業・ファイナンスなどに重要な応用がある.
- 統計的学習(Statistical learning)は現代のデータサイエンティスト(*data scientist*)の教育には基礎的で不可欠となっている.

教師なし学習
-Unsupervised learning-

- 目的変数は存在しない, あるサンプル集合上で予測量(特徴量)が与えられる.
- 目的は曖昧(more fuzzy) — サンプルから類似するグループを見つける, 類似の動きをする特徴量を見つける, 最大の変動を伴う特徴量の線形結合を見つけるなど.
- どの程度に上手く行っているのか知ることは困難.
- 教師あり学習(supervised learning)と異なるが、教師あり学習を行うための準備段階で有用となり得る.

ネットフリックス賞
-Netflix prize-

- 2006年10月に開始.トレーニングデータは18, 000映画のレーティング(400, 000 Netflix顧客,1-5).
- トレーニングデータは非常にスパース(very sparse)— 約98% は欠落データ.
- 目標は百万の顧客-映画のペアー(トレーニングデータでは欠落)のレーティングの予想.
- Netflix'sの原アルゴリズムではルートMSEは 0.953. この数値を10%改善できた最初のチームは百万ドル(1 million dollars)貰える.
- これは教師あり学習、あるいは教師なし学習のどちらかだろうか？

NETFLIX

Netflix Prize

COMPLETED

Home | Rules | Leaderboard | Update

Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

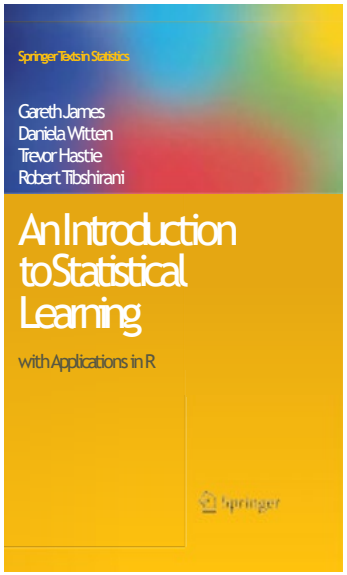
Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11

BellKor's Pragmatic Chaosチームが勝利, わずかにThe Ensembleチームを負かした.

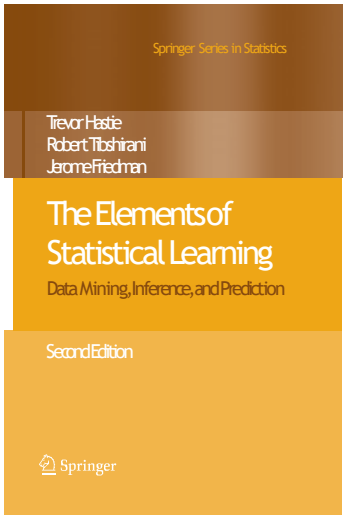
統計的学習と機械学習
-Statistical Learning vs. Machine Learning-

- 機械学習(Machine learning)は人工知能(Artificial Intelligence)の一分野として発展.
- 統計的学習(Statistical learning)は統計学(Statistics)の一分野として発展.
- *内容に重なりが多い* — 両者ともに教師あり, 教師なし学習問題に焦点をあてている:
  - 機械学習では大規模な応用や予測の精度*(prediction accuracy)*を強調.
  - 統計的学習ではモデル*(models)*, 解釈, 精度*(precision)*, 不確実性*(uncertainty)*を強調.
- しかし区別はまずますぼやけて来ていて、分野は交配 “cross-fertilization”している.
- 機械学習は特にマーケティング分野 *(Marketing!)*では標準になっている.

教科書



この講義はスプリンガー社から2021年に出版された教科書(ISLR第2版，左の写真の初版は2013年に出版)の内容をほぼカバーしている。本の各章にはR-labと実例が説明されている。書籍(初版，第2版)は著者のwebsitesからダウンロード可能である。

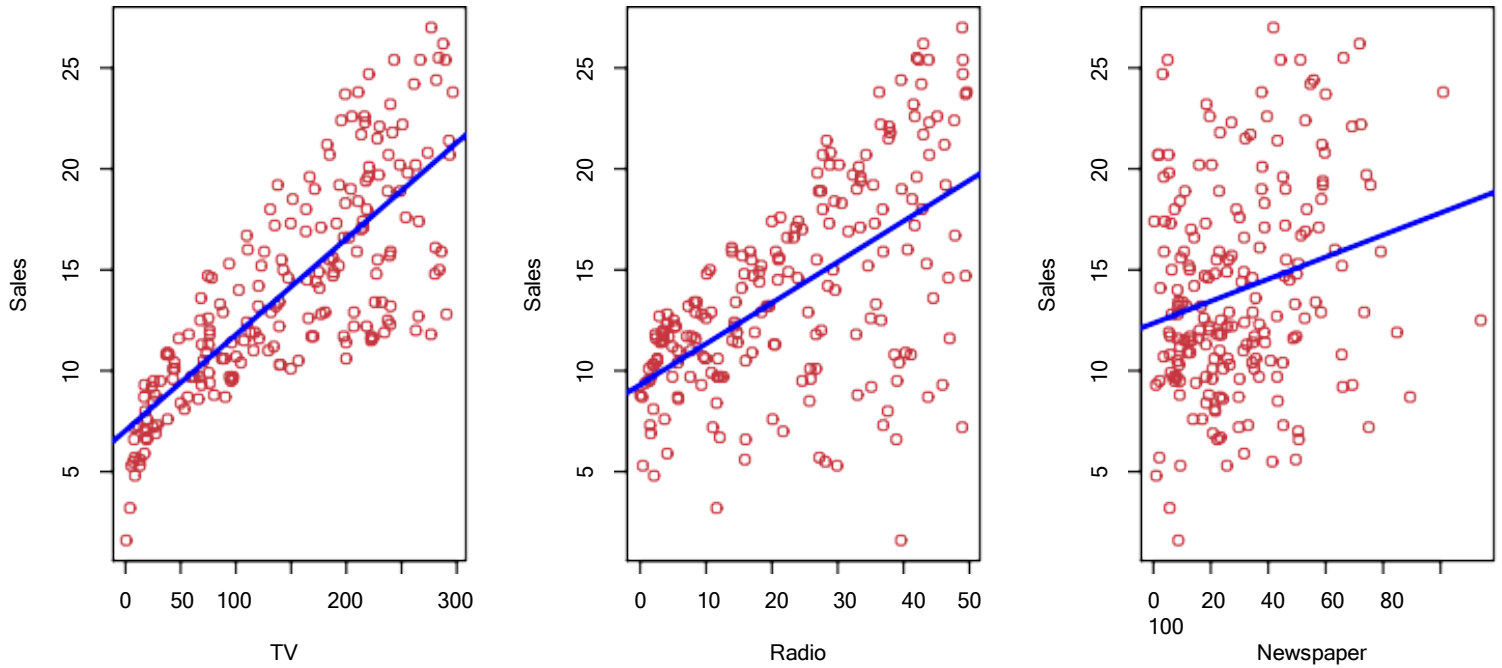


もう一つのスプリンガーから出版している書籍(ESL)はISLRに比べてより数理的により上級であり、2009年に第2版(共著者Jerome Friedman)が出版されている。取り上げられる話題はより広範囲である。書籍はスプリンガー社(Springer) アマゾン社(Amazon)から購入できるが、無料の電子版が著者のwebsiteからダウンロードできる。

第2章：統計的学習とは? -Statistical Learning-

- 売上とテレビ・ラジオ・新聞の広告
- 回帰関数-regression function-
- 最近傍平均
- パラメトリック・構造モデル
- 予測精度vs.解釈可能性
- モデル正確性の評価
- バイアス・分散のトレードオフ
- 分類問題-classification-

第2章：統計的学習とは?
-Statistical Learning-



売上(Sales)対テレビ(TV),ラジオ(Radio),新聞(Newspaper)
青の線形回帰直線はそれぞれのデータにフィットして得られた.
3つの直線を用いて売上(Sales)の予測ができるだろうか？
多分，次のモデルを用いると良い予測ができるだろう:

$$\text{売上(Sales)} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

記号(Notation)

ここで売り上げ(Sales)は予測したい反応変数(*response*),目的変数(*target*)である。ここで反応変数をYとする．テレビ(TV)は特徴量(*feature*), 入力変数,予測変数; 記号X_1で表そう．同様にラジオ(Radio)をX_2などとする．まとめると入力ベクトルにより次のように表現する:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

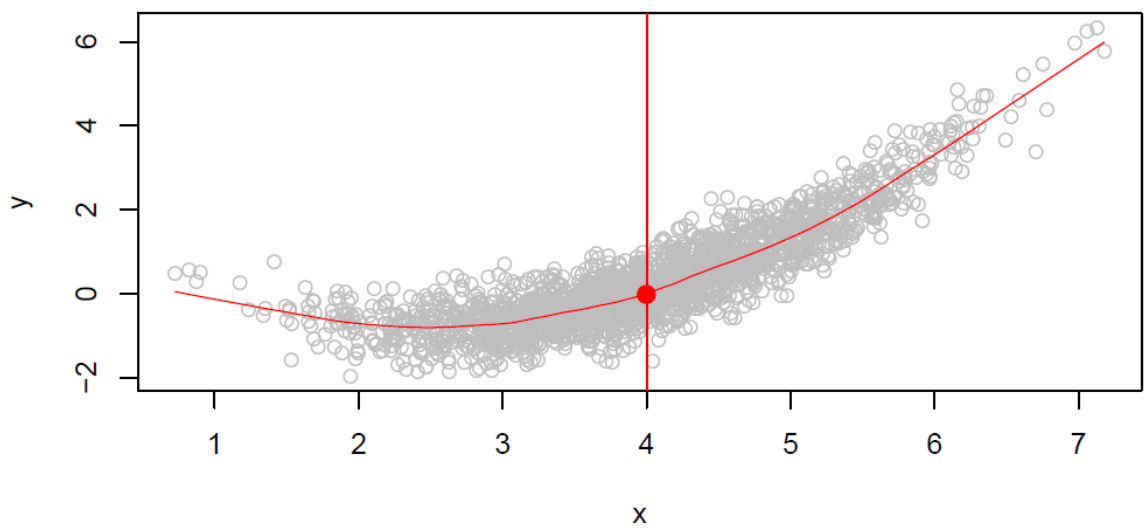
ここで次の統計モデルを扱う:

$$Y = f(X) + \epsilon$$

ただしϵは計測誤差(measurement errors)および他の誤差(discrepancies)を示している.

どの様な$f(X)$が適切か？

- 適切な f を用いると，新たな点$X = x$においてYの予測が可能となる．
- 変数 $X = (X_1, X_2, \ldots, X_p)$ のどの要素が変数Yを説明するのに重要な役割を果たす，あるいは不適切なのか理解できる．例えば経験(Seniority)と教育年数(Years of Education)は所得(Income)に大きな影響があるが，結婚・既婚(Marital Status)は普通の場合は影響しない．
- 関数 f の複雑性にも依存するが，Xの各要素 X_j がどの様にYに影響するか理解できる．



理想的な $f(X)$が考えられるだろうか? 特に任意のXの値,
$X = 4$に対して適切な値 $f(X)$は何だろうか？
$X = 4$に対して様々なYが考えられる. ここで適切な値は

$$f(4) = E(Y|X = 4)$$

となる. ここで$E(Y|X = 4)$ は条件$X = 4$が与えられた時のYの平均,
期待値(*expected value,* 平均)を意味する.

この適切な $f(x) = E(Y|X = x)$は回帰関数(*regression function*)と
呼ばれている.

回帰関数(regression function) $f(x)$

- ベクトルXの場合も定義でき; 例えば

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

回帰関数(regression function) $f(x)$

- ベクトルXの場合も定義でき; 例えば
  $$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$
- すべての$X = x$の点に対するすべての関数gに対して $f(x) = E(Y|X = x)$ は平均二乗予測誤差(mean-squared prediction error) $E[(Y - g(X))^2|X = x]$ を最小化する : Yの理想的,あるいは最適な(*optimal*)予測量となる.

回帰関数(regression function) $f(x)$

- ベクトルXの場合も定義でき; 例えば
  $$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$
- すべての$X = x$の点に対するすべての関数gに対して $f(x) = E(Y|X = x)$ は平均二乗予測誤差(mean-squared prediction error) $E[(Y - g(X))^2|X = x]$ を最小化する : Yの理想的,あるいは最適な(*optimal*)予測量となる.
- $\epsilon = Y - f(x)$ は予測誤差(*irreducible* error) — すなわち,

仮に関数$f(x)$を知っていたとしても,予測する場合には誤差が生じる,
各$X = x$に対してYは通常はある分布にしたがうと考えられる.

回帰関数(regression function) $f(x)$

- ベクトルXの場合も定義でき; 例えば
$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$
- すべての$X = x$の点に対するすべての関数gに対して $f(x) = E(Y|X = x)$ は平均二乗予測誤差(mean-squared prediction error) $E[(Y - g(X))^2|X = x]$ を最小化する : Yの理想的,あるいは最適な(*optimal*)予測量となる.
- $\epsilon = Y - f(x)$ は予測誤差(*irreducible* error) — すなわち,

仮に関数$f(x)$を知っていたとしても,予測する場合には誤差が生じる,
各$X = x$に対してYは通常はある分布にしたがうと考えられる.

- 関数 $f(x)$の任意の推定量 $\hat{f}(x)$ に対して次の関係がある

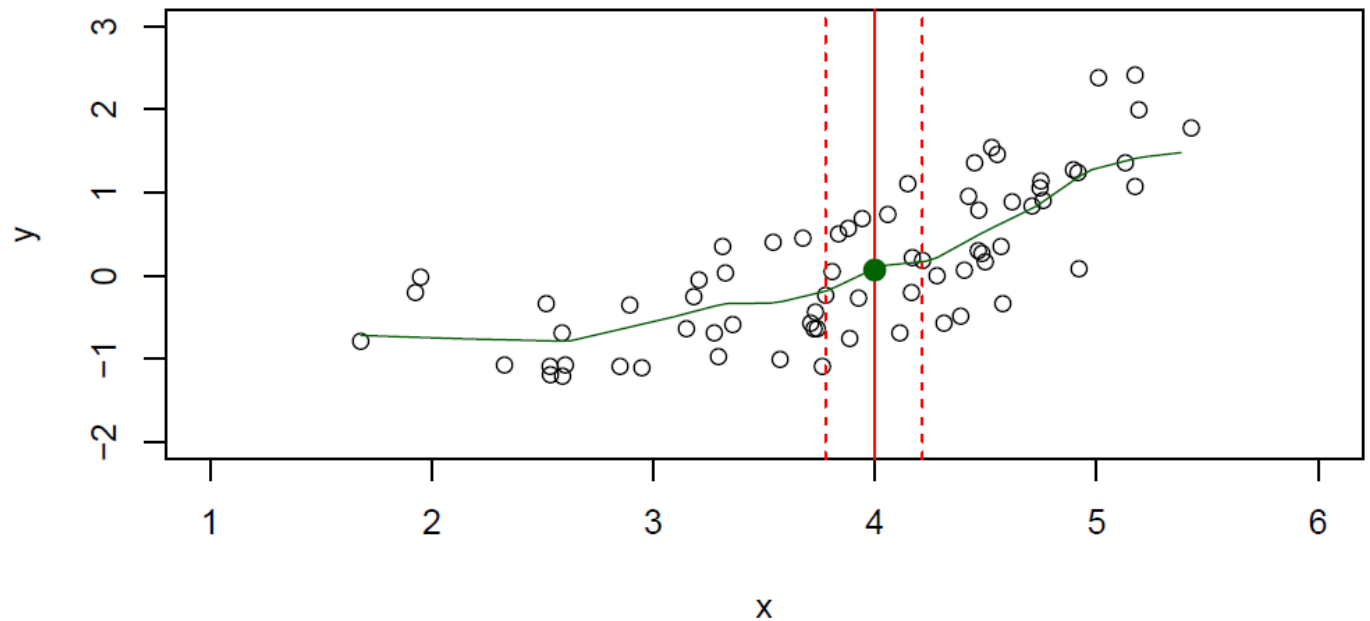
$$E\left[(Y - \hat{f}(X))^2 \middle| X = x\right] = \underbrace{\left[f(x) - \hat{f}(x)\right]^2}_{\text{改善可能項}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{改善不能項}}$$

関数 f の推定

- 典型的には正確に$X = 4$となるデータ点があったとしても多くない.
- したがってデータから $E(Y|X = x)$を正確に計算できない!
- そこで少し定義を緩めて推定値

$$\hat{f}(x) = \mathrm{Ave}(Y|X \in \mathcal{N}(x))$$

を考察しよう. ただし $\mathcal{N}(x)$ は点xの近傍(*neighborhood*)とする.

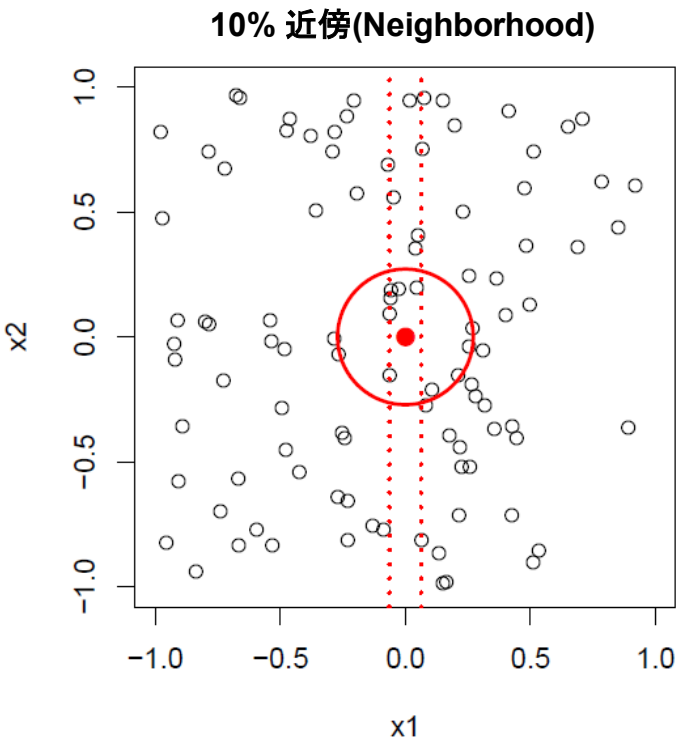


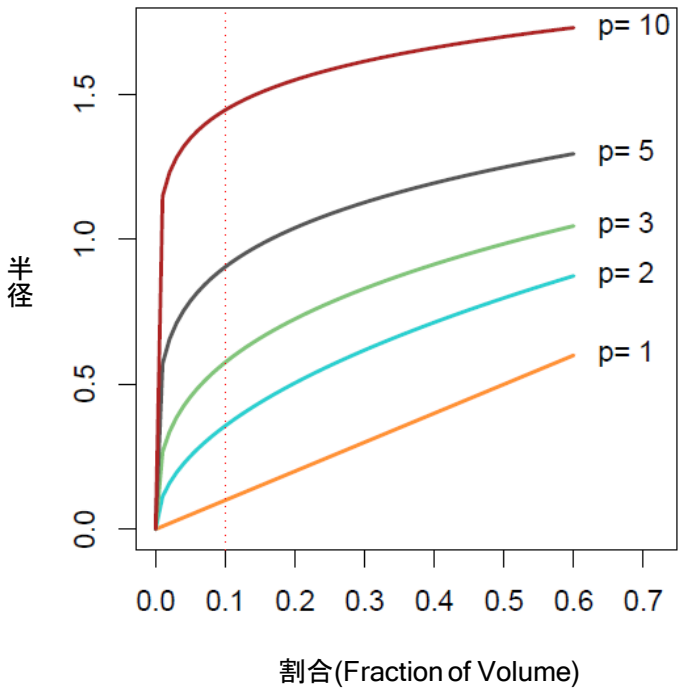
- 最近傍平均(Nearest neighbor averaging)はpが小さければよい性能を持っている— 特に$p \leq 4$,データ数N大きい場合.
- この講義では後でNNAの平滑化(smoother)法, 例えばカーネル平滑化, スプライン平滑化などを議論する.

- 最近傍平均(Nearest neighbor averaging)はpが小さければよい性能を持っている— 特に$p \leq 4$,データ数N大きい場合.
- この講義では後でNNAの平滑化(smoother)法, 例えばカーネル平滑化, スプライン平滑化などを議論する.
- 最近傍法はpが大きいとかなりお粗末(*lousy*)なりうる.
  理由: 次元の呪い(*curse of dimensionality*). 最近傍のデータは高次元の場合にはかなり遠くなる傾向がある.

    - N個の各y_iに対して適切な一部分, 例えば10%を使って平均化, 分散を小さくすることが必要となる.
    - 高次元では10% 近傍は局所的とは限らず, 局所的平均化して$E(Y|X = x)$ を推定するという考え方は適切とは限らない.

次元の呪い(curse of dimensionality)





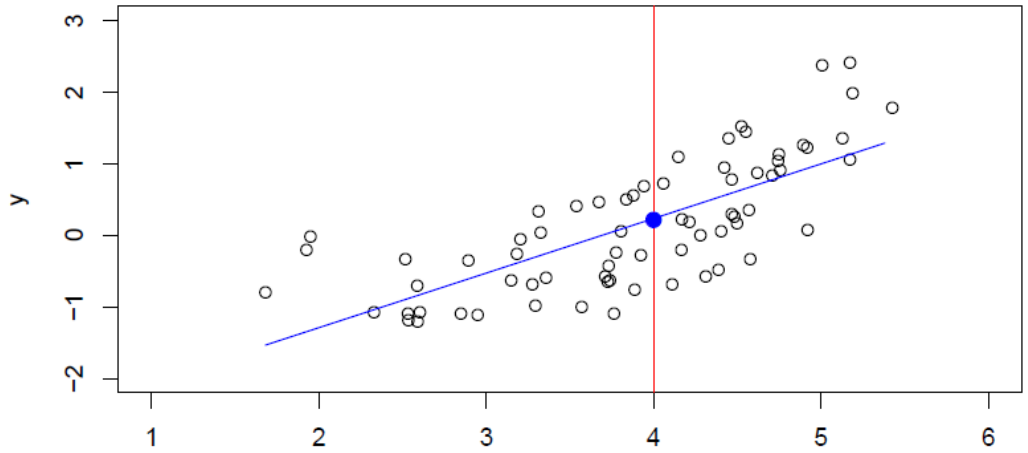
パラメトリック・構造モデル
(Parametric and structured models)

線形モデル(*linear* model)はパラメトリックモデル(parametric model)の重要な例:

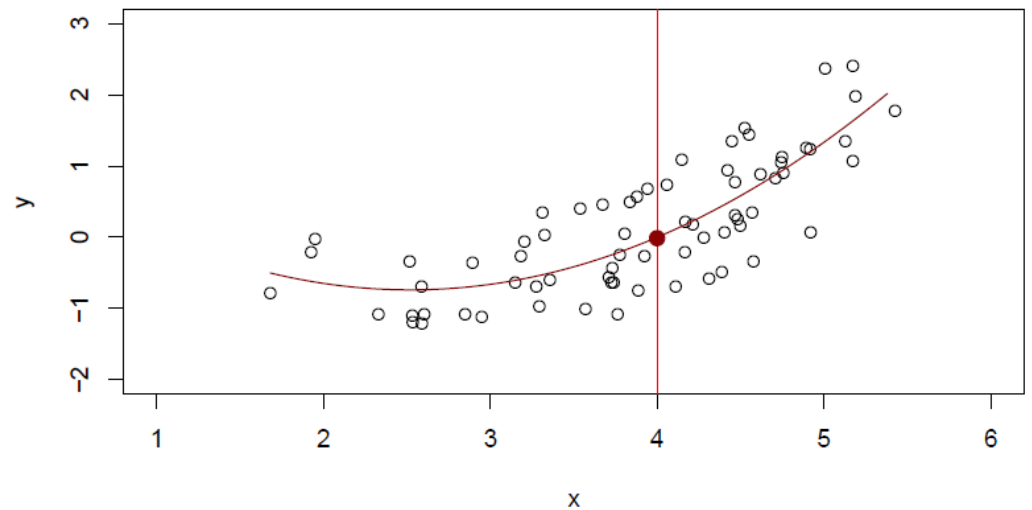
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

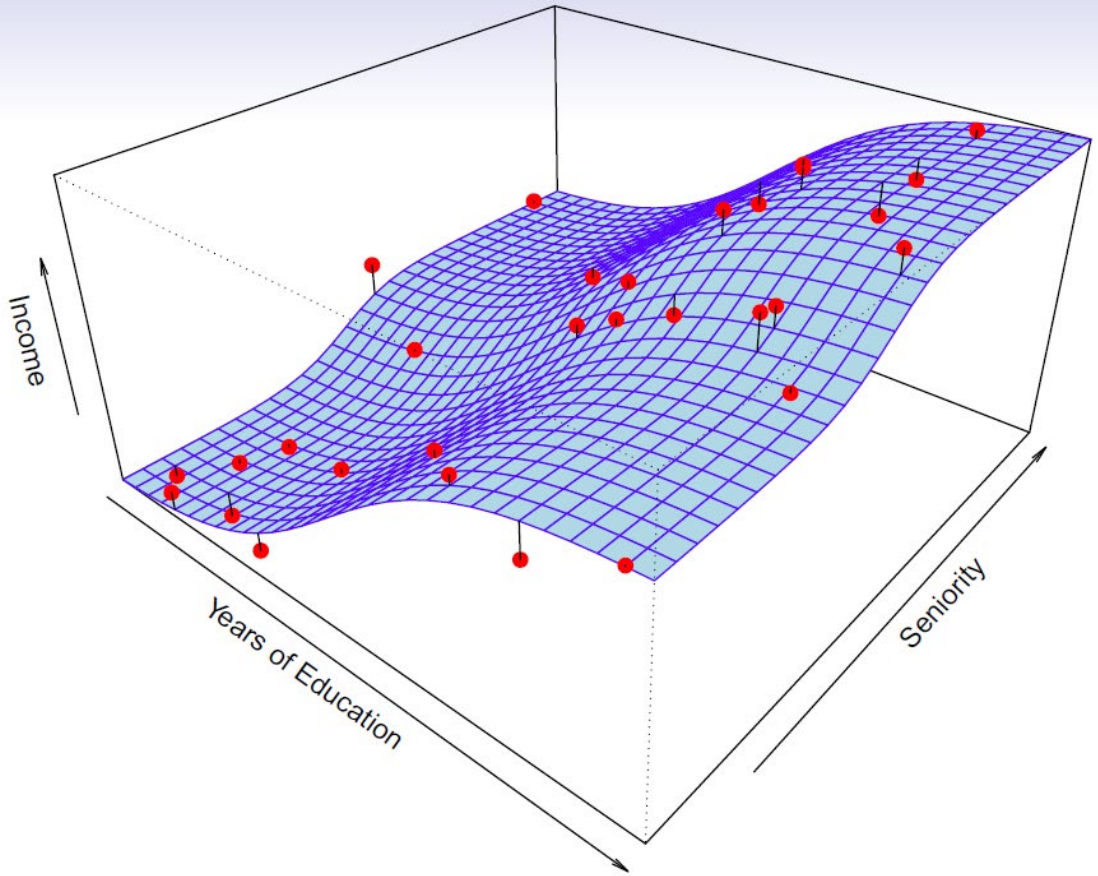
- 線形モデルは$p + 1$個のパラメター(母数,parameters) $\beta_0, \beta_1, \ldots, \beta_p$で定める.
- モデルを訓練データ(training data)にフィットすることによりパラメターを推定する.
- 線形モデルはほとんどの場合は正しくない(*almost never correct)* ではあるがしばしば未知の真の関数$f(X)$に対する適切, あるいは解釈可能な近似として役立つ.

線形モデル $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ はこの例では良くフィットしている



2次モデル $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ の方がフィットは少し良い.

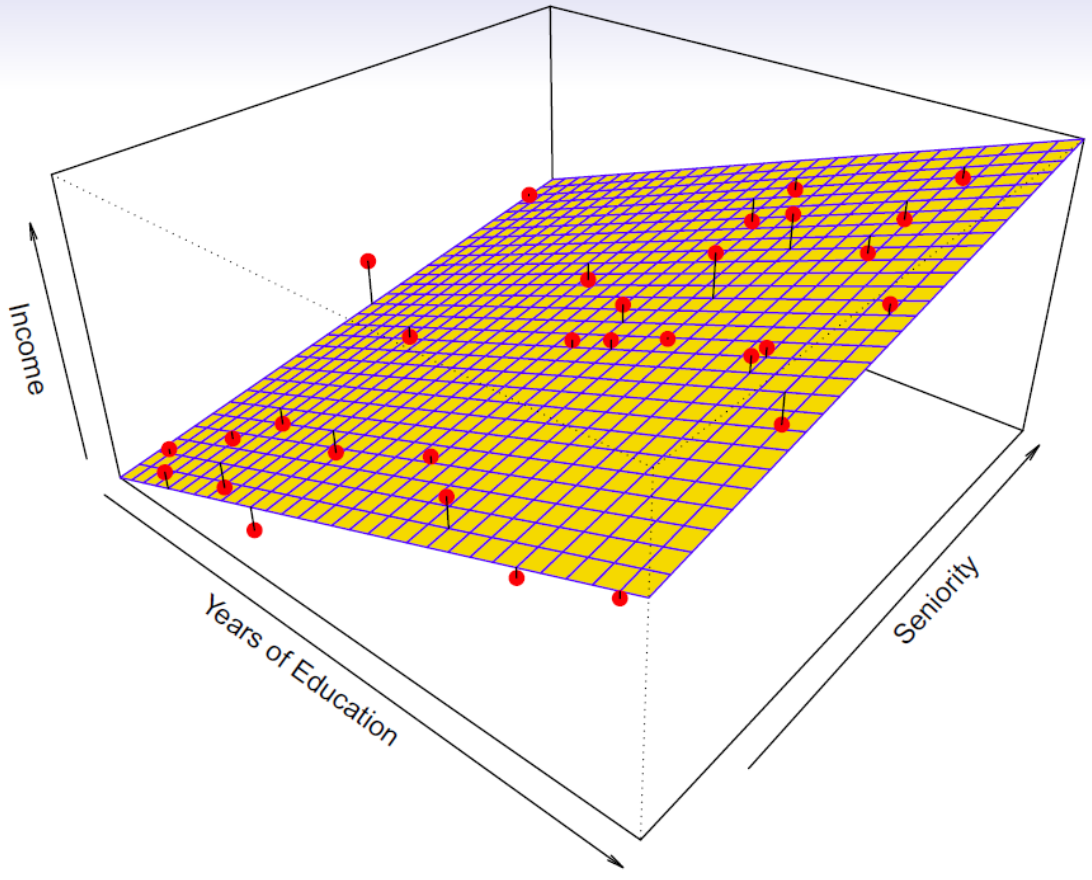




シミュレーションの例. 赤の点Red pointsは次のモデルから所得変数をシミュレートした値

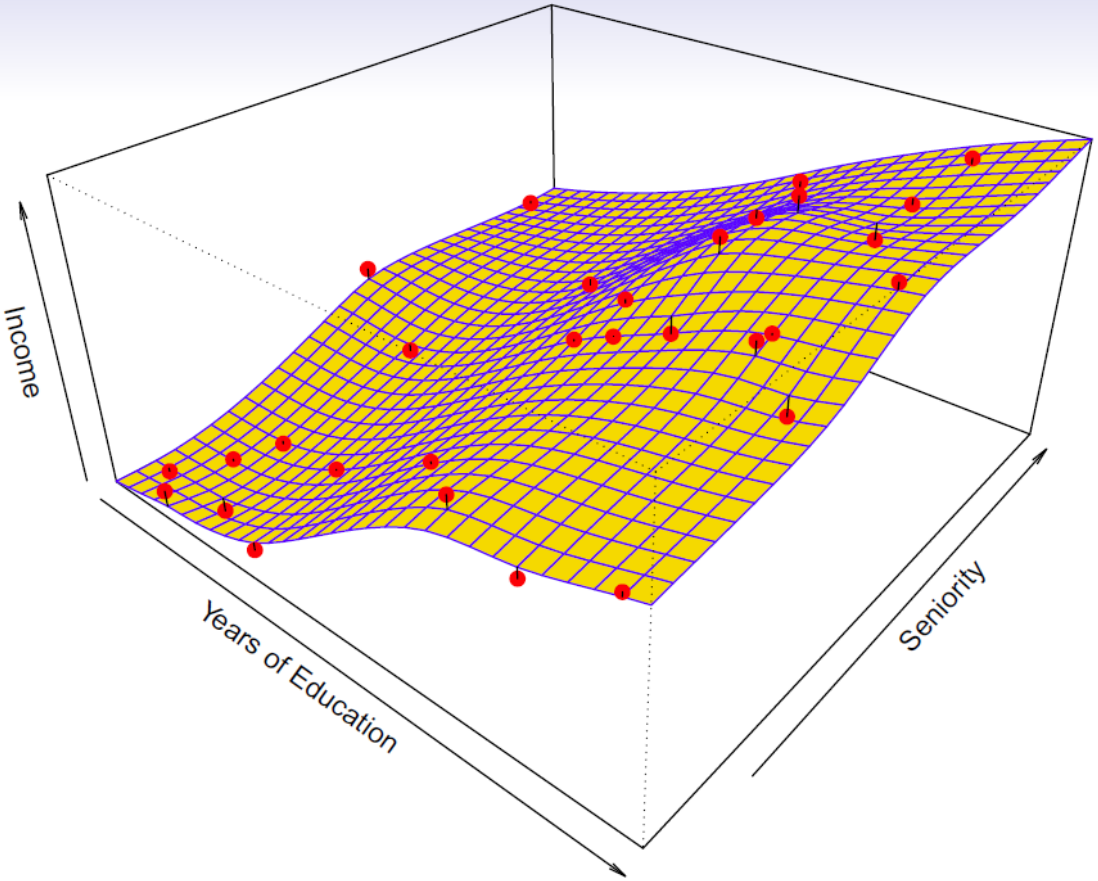
$$\text{所得(income)} = f(\text{教育(education)}, \text{経験年数(seniority)}) + \epsilon$$

f は青の曲面である.

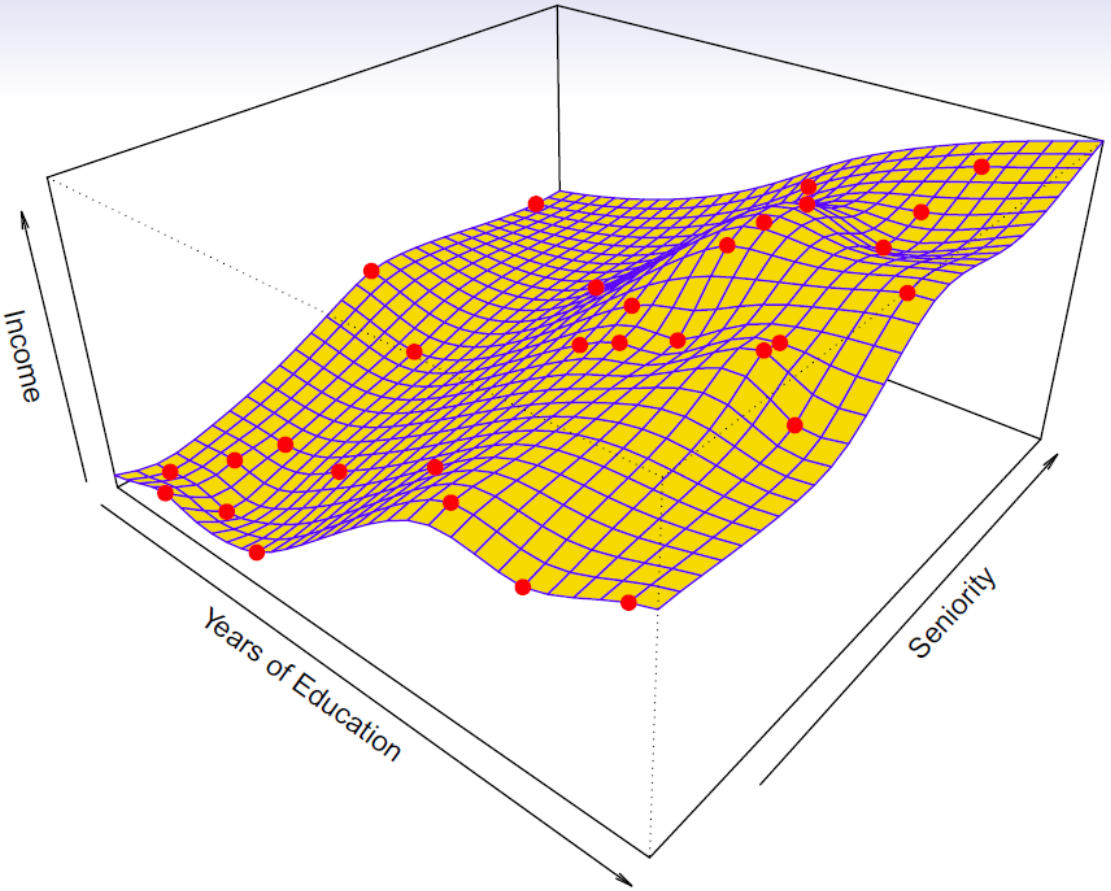


線形モデルをシミュレーションデータにフィットすると：

$\hat{f}_L(\text{education}, \text{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \text{education} + \hat{\beta}_2 \times \text{seniority}$



より柔軟な統計モデル $\hat{f}_S$(教育(education), 経験年数(seniority)) をシミュレーション・データにフィットする. ここでは統計的方法である薄版スプライン(*thin-plate spline*)により柔軟な曲面をフィットしてみよう. フィットの滑らかさを制御できる (7章を参照).



さらにより柔軟なスプライン回帰モデル

$\hat{f}_S$(教育(education), 経験年数(seniority)) をフィットすることもできる.

この場合には訓練データの上ではフィットしたモデルでは誤差はゼロとなっている! これは過適合(*overfitting*)の例だろう.

トレードオフ
(trade-offs)

- 予測精度vs.解釈可能性(Prediction accuracy vs. interpretability).
  — 線形モデルは解釈しやすいが, 薄板スプライン(thin-plate splines)モデルはそうでもない.

トレードオフ
(trade-offs)

- 予測精度vs.解釈可能性(Prediction accuracy vs. interpretability).
  — 線形モデルは解釈しやすいが, 薄板スプライン(thin-plate splines)モデルはそうでもない.

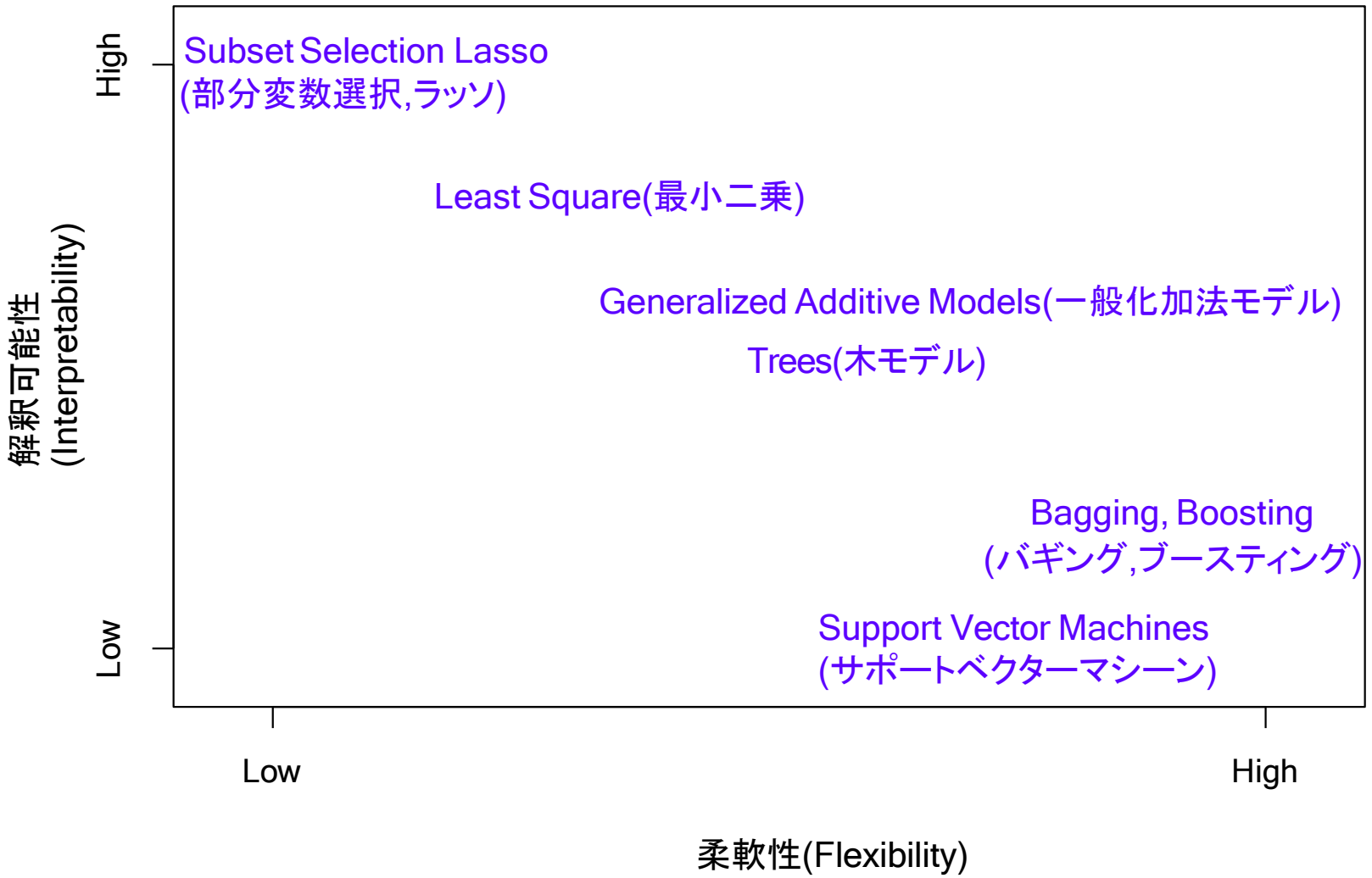
- 良いフィットvs. 過適合(over-fit), 過少適合(under-fit).
  — どうしたらフィットが適切か否かがわかるのだろうか？

トレードオフ
(trade-offs)

- 予測精度vs.解釈可能性(Prediction accuracy vs. interpretability).
  — 線形モデルは解釈しやすいが, 薄板スプライン(thin-plate splines)モデルはそうでもない.

- 良いフィットvs. 過適合(over-fit), 過少適合(under-fit).
  — どうしたらフィットが適切か否かがわかるのだろうか？

- ケチの原理(parsimony) vs. ブラックボックス(black-box).
  — しばしば統計家はすべての変数を含むブラックボックスモデルよりも, 少ない変数を含んでいる単純な統計モデルを好む.


Subset Selection Lasso
(部分変数選択,ラッソ)
Least Square(最小二乗)
Generalized Additive Models(一般化加法モデル)
Trees(木モデル)
Bagging, Boosting
(バギング,ブースティング)
Support Vector Machines
(サポートベクターマシーン)
High
Low
解釈可能性
(Interpretability)
Low
High
柔軟性(Flexibility)

モデル正確性の評価
(Model Accuracy)

ここで統計モデル $\hat{f}(x)$を訓練データ
Tr $= \{x_i, y_i\}_1^N$にフィット, モデルの良さを評価したいとする.

- データ Tr 上で平均二乗予測誤差を計算してみることができる:

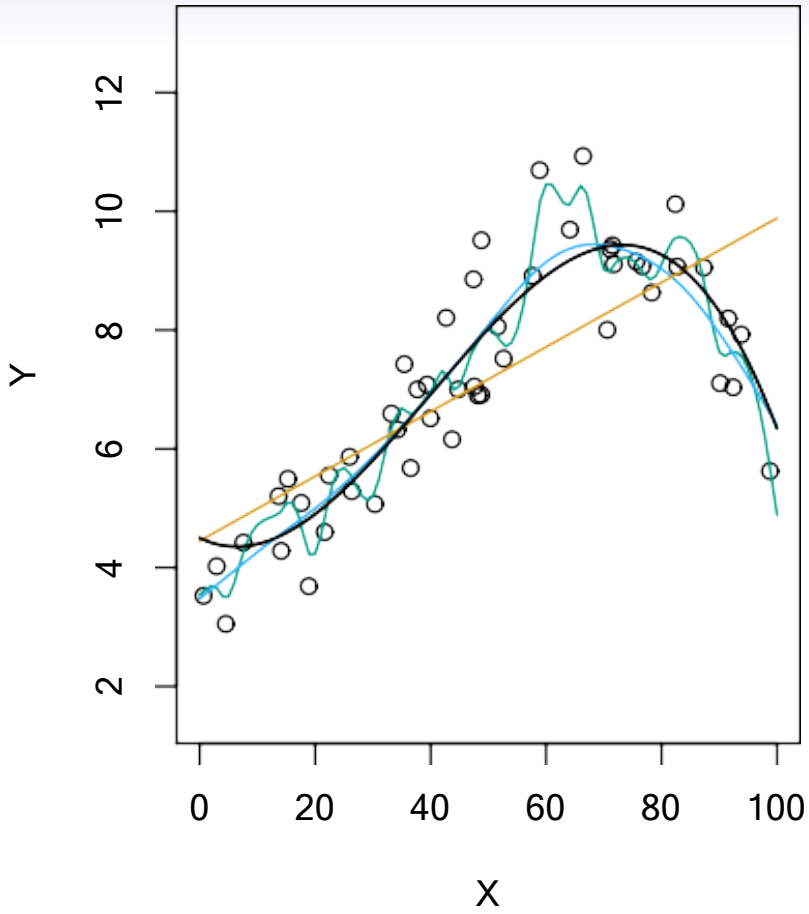
$$\mathrm{MSE}_{\mathrm{Tr}} = \mathrm{Ave}_{i \in \mathrm{Tr}} \left[y_i - \hat{f}(x_i) \right]^2$$

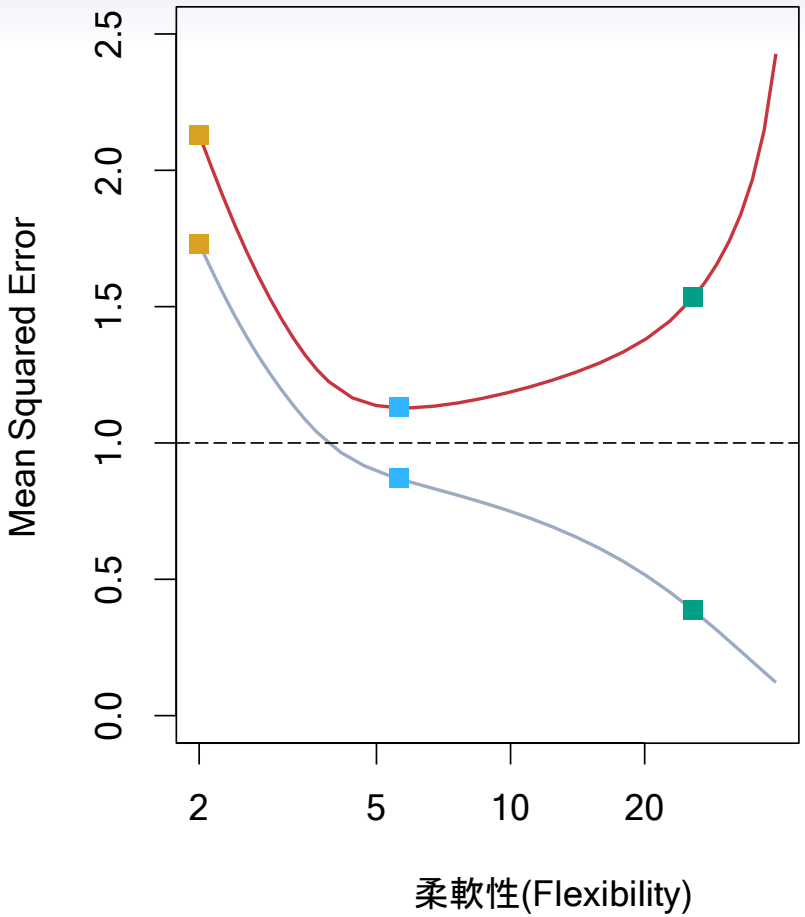
この評価はモデルの過適合になるバイアスがある.

- この場合には可能であれば新たなテスト(*test*)データを使って平均二乗予測誤差を求めるべきである.

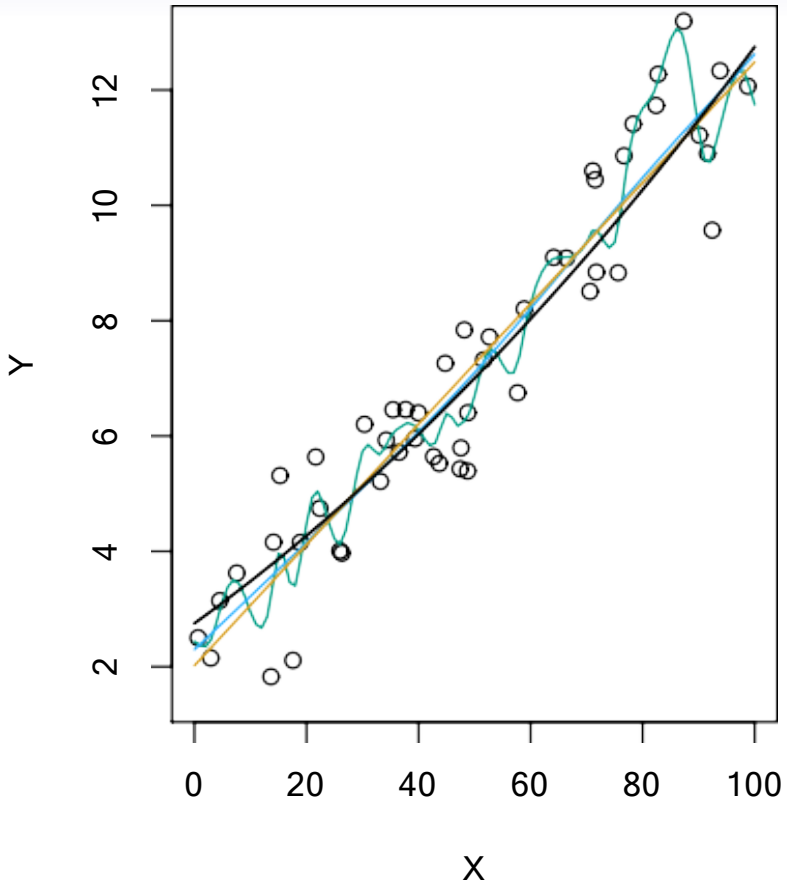
データ Te$= \{x_i, y_i\}_1^M$:

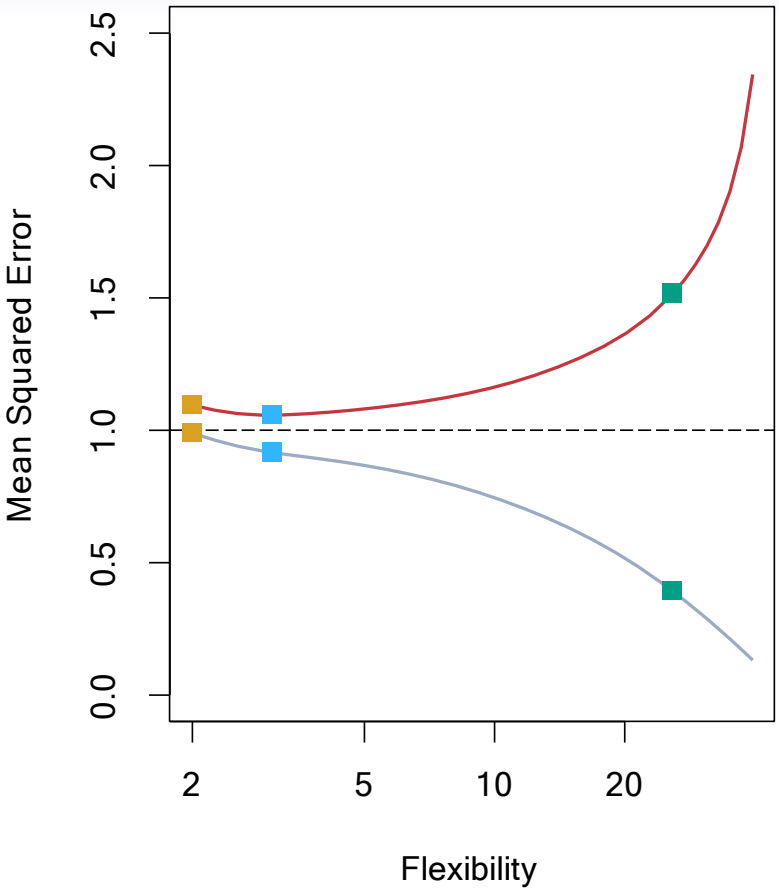
$$\mathrm{MSE}_{\mathrm{Te}} = \mathrm{Ave}_{i \in \mathrm{Te}} \left[y_i - \hat{f}(x_i) \right]^2$$



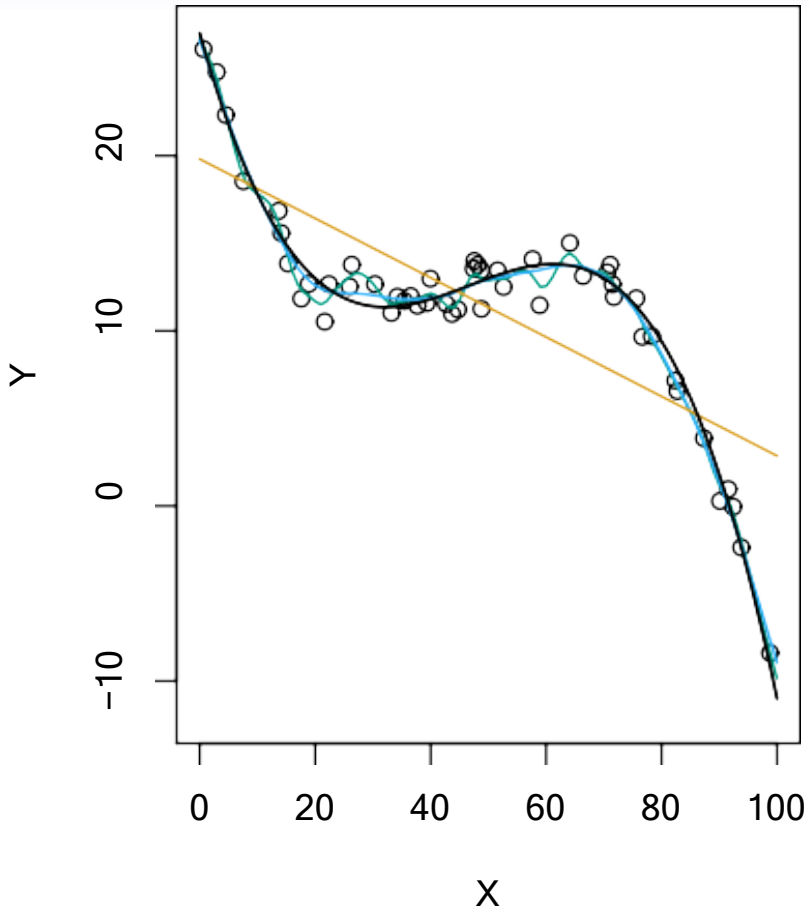


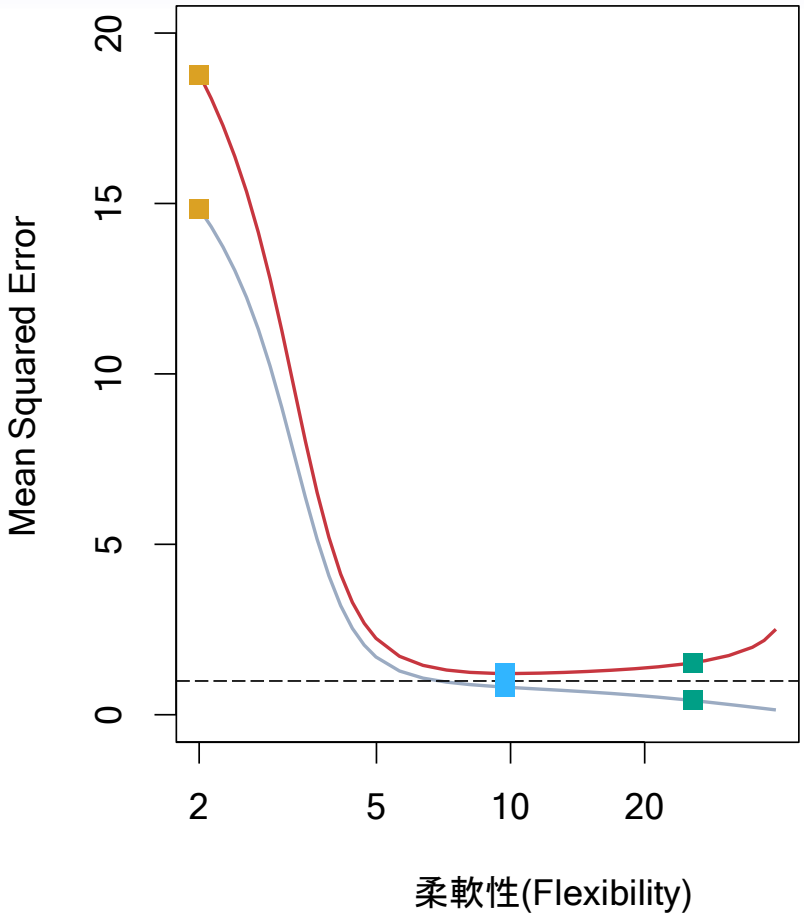
黒の曲線が真. 右図の赤の曲線が $\mathrm{MSE}_{\mathrm{Te}}$ 最小, グレイの曲線が $\mathrm{MSE}_{\mathrm{Tr}}$ を最小化. オレンジ, 青, グリーンの曲線はそれぞれ柔軟なフィットに対応している.





上の例では真の曲線はより滑らかなのでフィットした曲線はより滑らかとなり線形モデルはかなり良くなっている.





真の形状は波状, ノイズは小さい, したがって柔軟な曲線フィットが最適になる.

バイアス・分散のトレードオフ (Bias-Variance Trade-off)

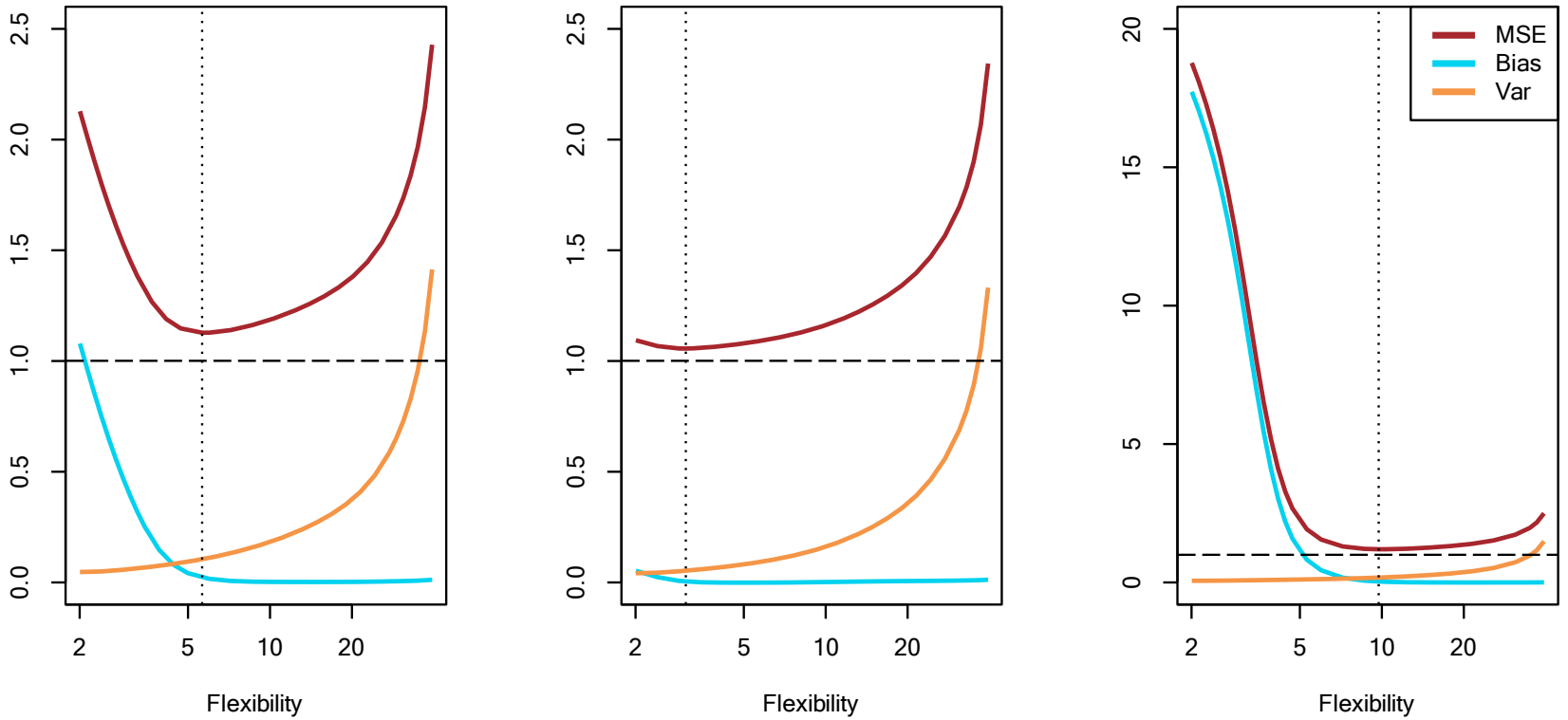
訓練データ Tr にある統計モデル $\hat{f}(x)$をフィット, (x_0, y_0)を母集団から抽出したテスト観測値としよう. もし真の統計モデルが $Y = f(X) + \epsilon$ (ただし$f(x) = E(Y|X = x)$) とすると,

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

ここで期待値は Tr における変動性に加えてy_0の変動性を含んで平均化している. バイアスは$\mathrm{Bias}(\hat{f}(x_0)) = E\left[\hat{f}(x_0)\right] - f(x_0)$ である.

典型的には $\hat{f}$の柔軟性(*flexibility*)が増せば, 分散が増加, バイアスが減少する. したがって平均テスト誤差にもとづいて柔軟性を選択するとバイアス・分散のトレードオフ(*bias-variance trade-off*)に対処できる.

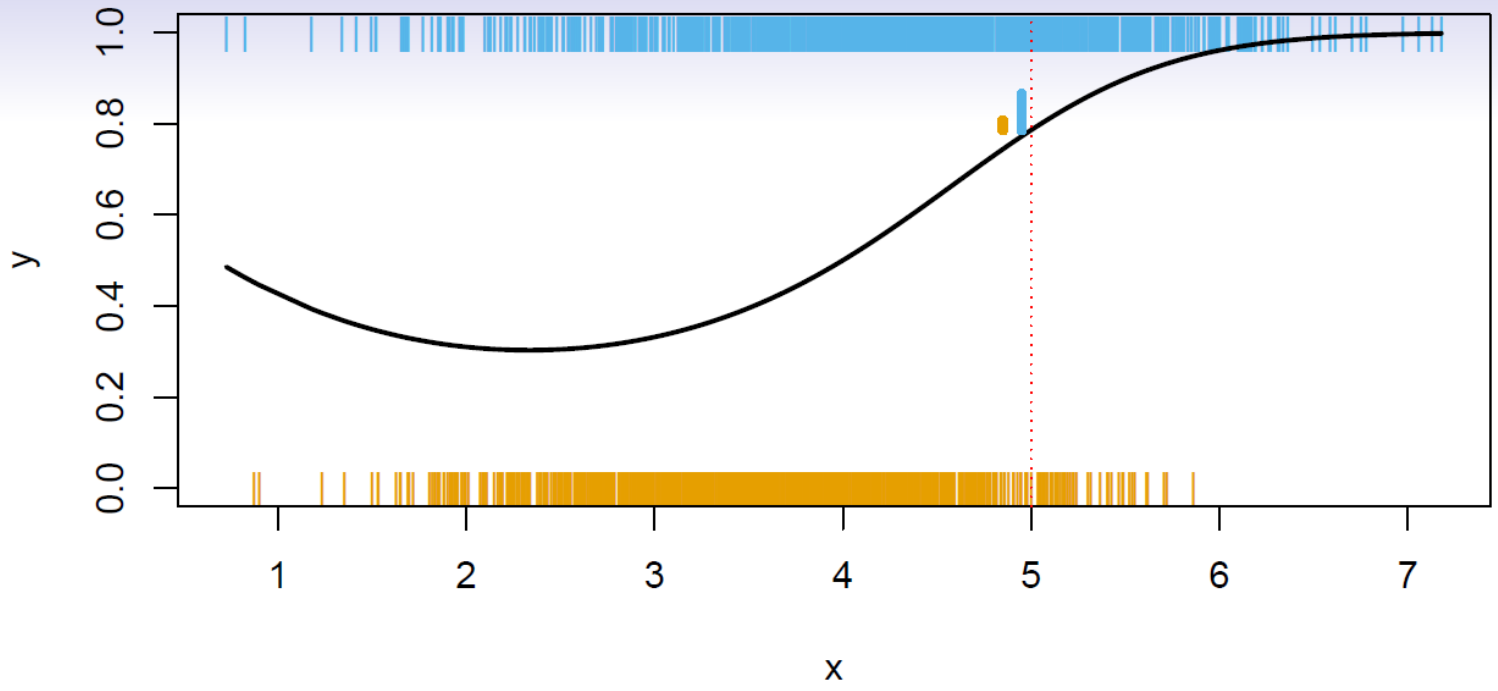
バイアス・分散のトレードオフの3つの例
(Bias-variance trade-off)



分類問題
(Classification)

反応変数(response variable) Yが質的変数(*qualitative*)な場合 — 例えばeメールが$\mathcal{C} = \{$スパムspam, 無害ham$\}$ (ham=無害good email)}のいずれかの場合, 10個のクラスが $\mathcal{C} = \{0, 1, \ldots, 9\}$のいずれかの場合など. ここでの目的は:

- ある分類器(classifier) $C(X)$を作成, 得られるラベルなしの観測値Xに$\mathcal{C}$からあるクラスを割り付ける.
- 各分類における不確実性の評価.
- 異なる予測量$X = (X_1, X_2, \ldots, X_p)$についての役割を理解する.



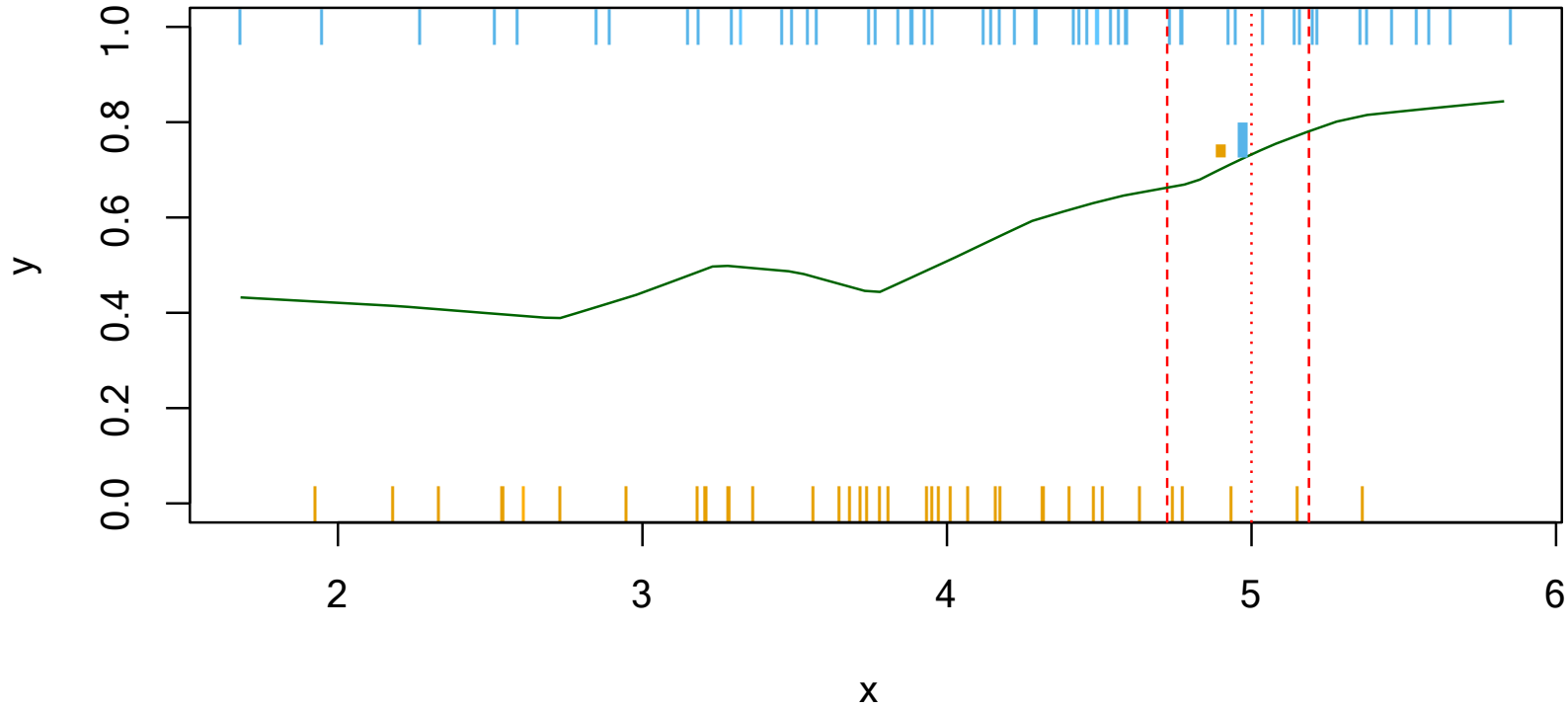
理想的な$C(X)$が存在するだろうか？ 仮に$\mathcal{C}$にK個の要素$1,2,\ldots,K$が割り付けられたとする. ここで

$$p_k(x) = \Pr(Y = k|X = x), k = 1,2,\ldots,K$$

とすると, xにおける条件付確率(*conditional class probabilities*)

である; つまり$x = 5$の棒グラフを見よう. このときxにおけるベイズ最適分類器(*Bayes optimal* classifier)は

$$C(x) = j \text{ if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$



最近傍平均(Nearest-neighbor averaging)を前と同様に利用できる.
しかし次元が大きくなるとあまり上手くいかない. ただし$\hat{C}(x)$ への
影響は$\hat{p}_k(x)\ k = 1, \ldots, K$ に対するよりも小さくなる.

分類(Classification): 詳 細

- 普通は分類器 $\hat{C}(x)$ のパフォーマンスは誤分類率 (misclassification error rate)により評価される:

$$\mathrm{Err}_{\mathrm{Te}} = \mathrm{Ave}_{i \in \mathrm{Te}} I[y_i \neq \hat{C}(x_i)]$$

- ベイズ分類器 (ただし正しい$p_k(x)$ を用いた場合) は(母集団では)最小の誤差を持つ.

分類(Classification): 詳 細

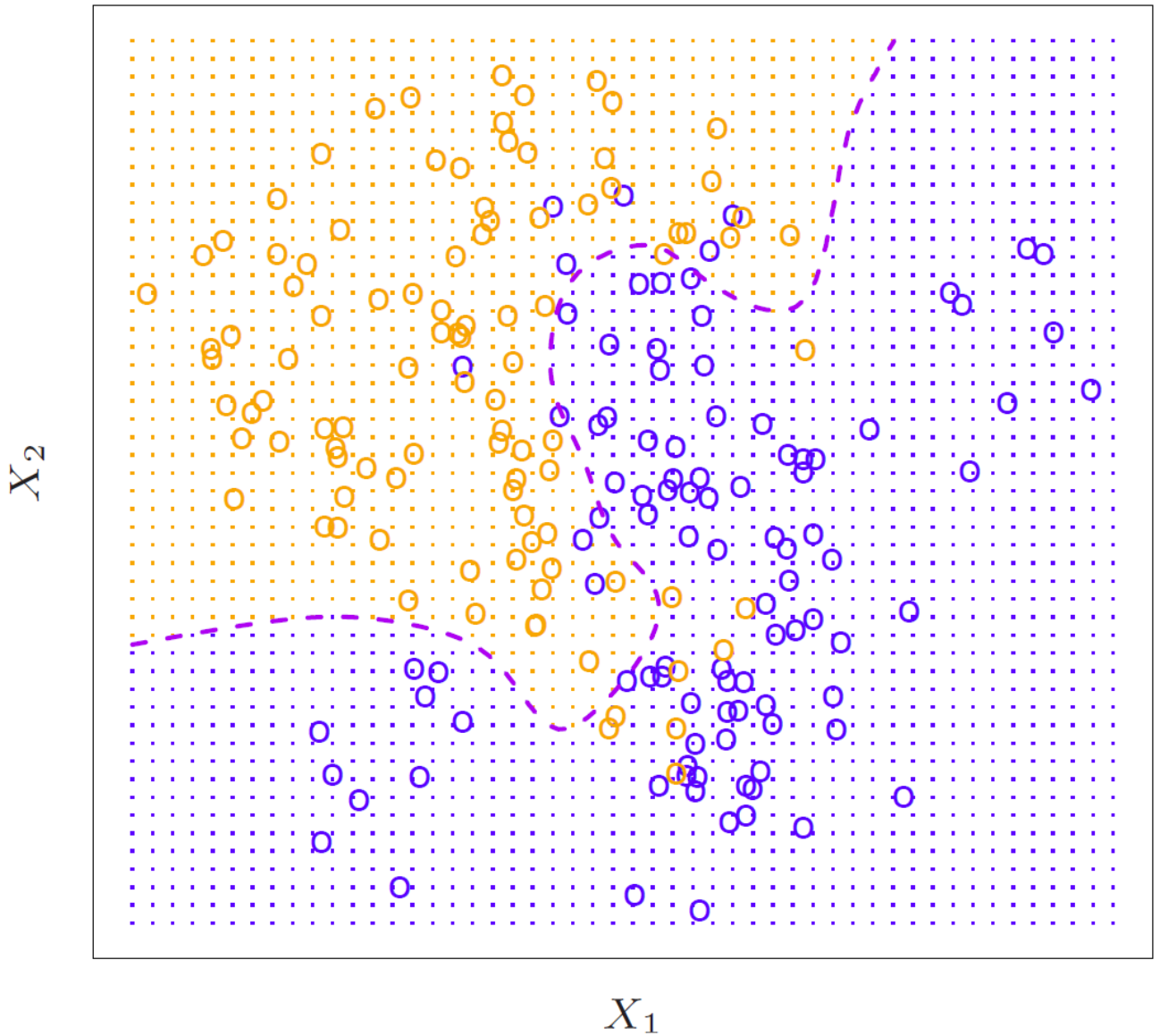
- 普通は分類器 $\hat{C}(x)$ のパフォーマンスは誤分類率 (misclassification error rate)により評価される:

$$\mathrm{Err}_{\mathrm{Te}} = \mathrm{Ave}_{i \in \mathrm{Te}} I[y_i \neq \hat{C}(x_i)]$$

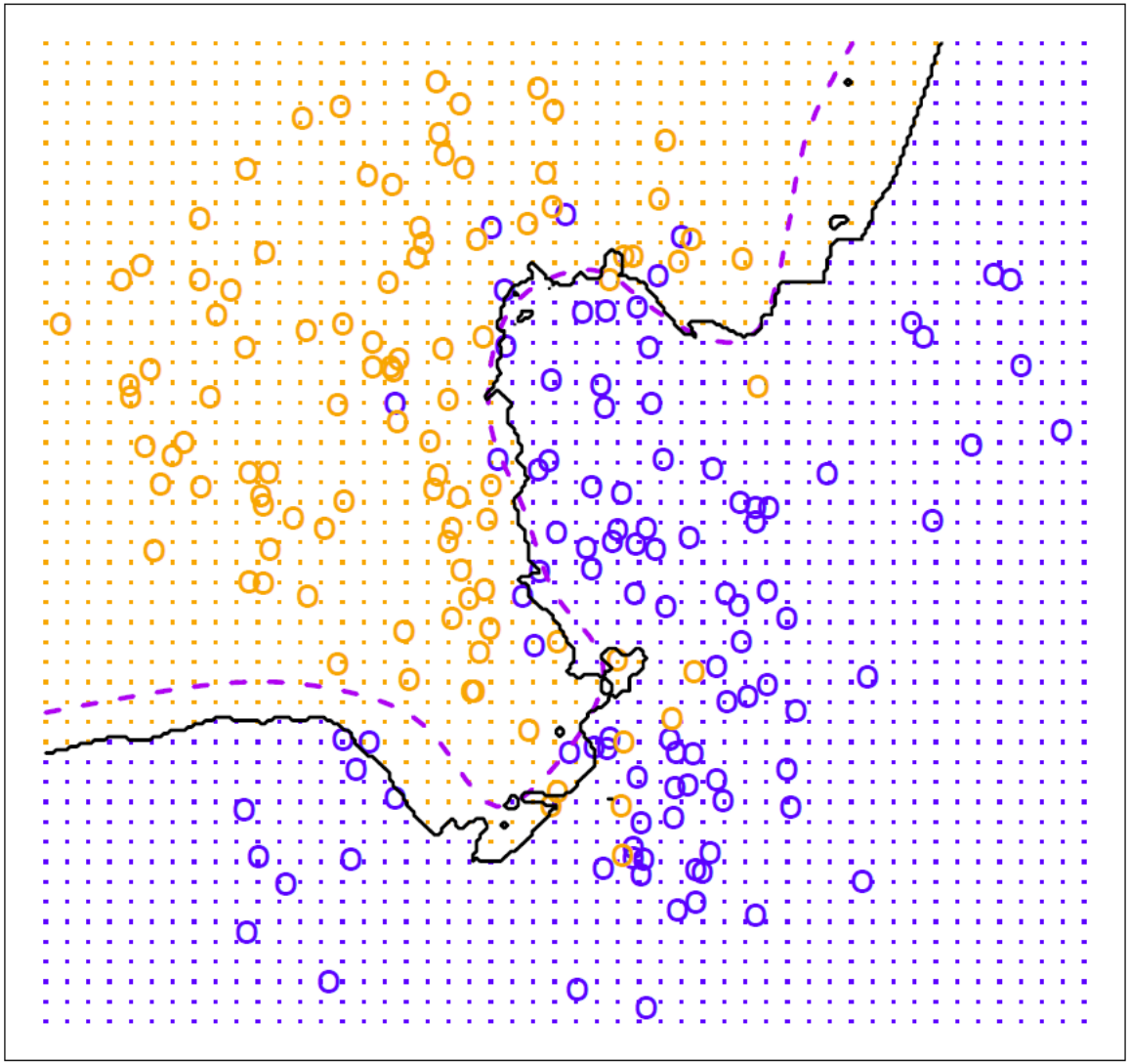
- ベイズ分類器 (ただし正しい$p_k(x)$ を用いた場合) は(母集団では)最小の誤差を持つ.

- サポートベクター分類器(Support-vector machines)を$C(x)$に対する統計モデルを作ることができる.
- また確率$p_k(x)$を表現する統計モデル(例えばロジッティック回帰, 一般化加法モデル など)を作ることもできる.

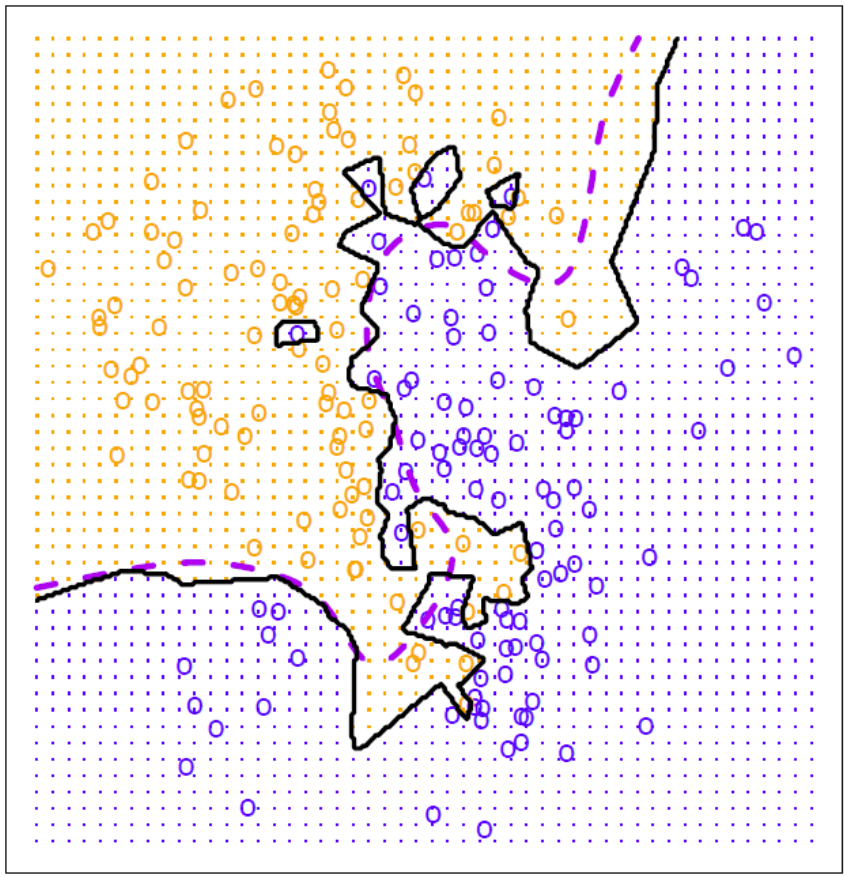
例: 2次元のK-最近傍法(nearest neighbors)



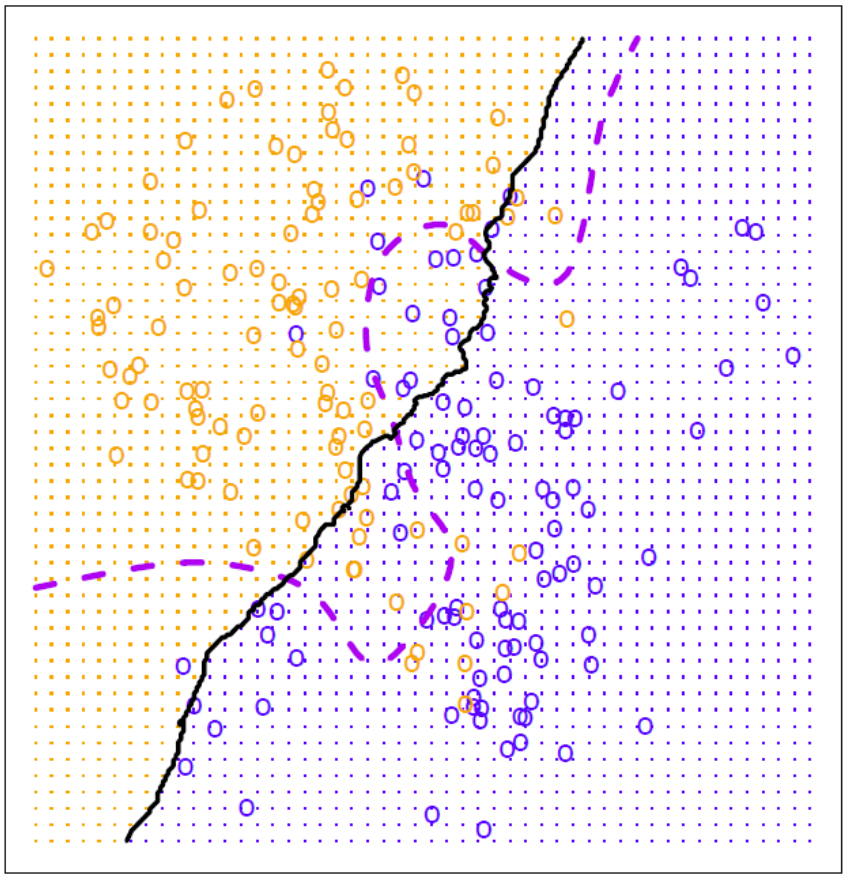
KNN: K=10

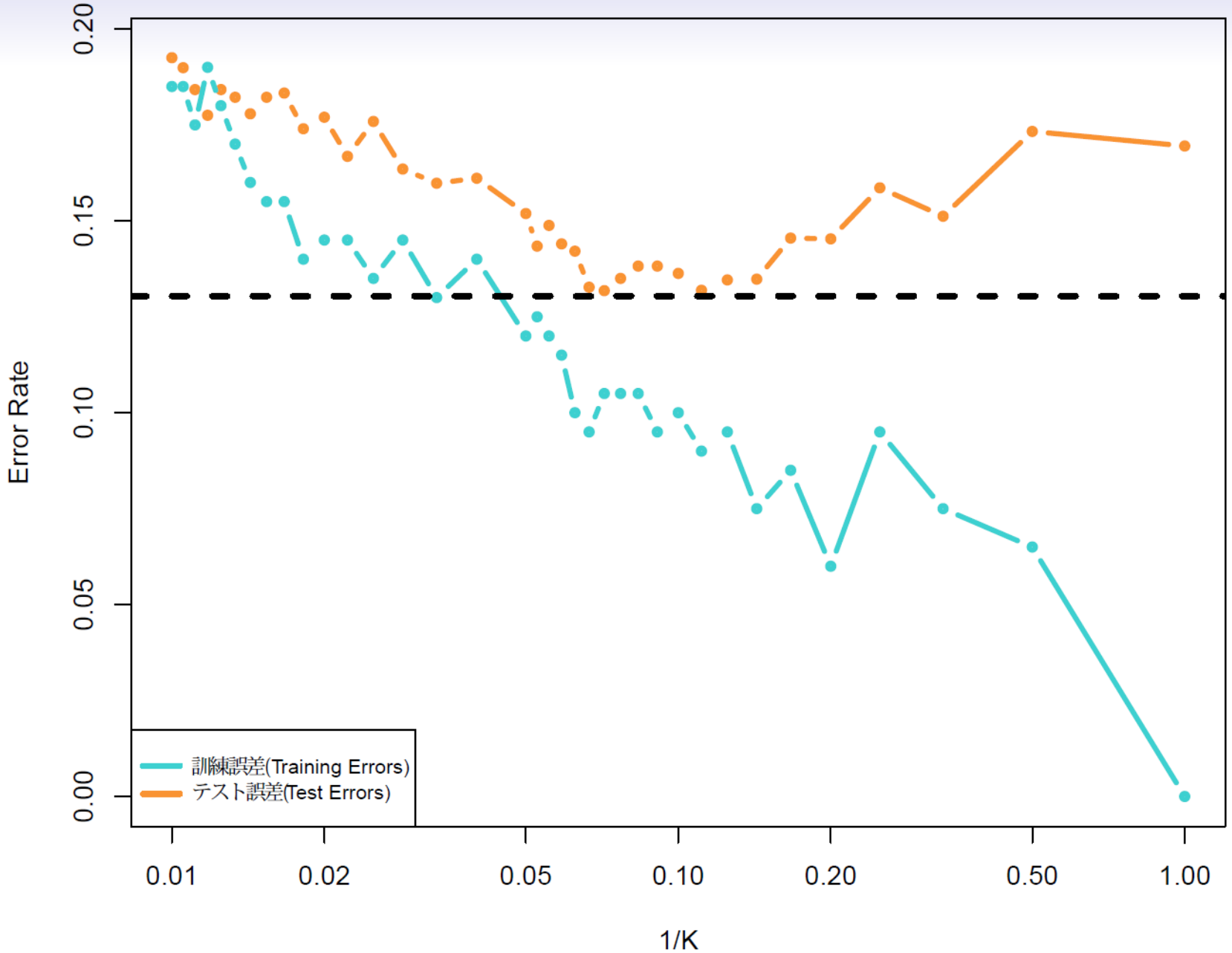


KNN: K=1



KNN: K=100




Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
訓練誤差(Training Errors)
テスト誤差(Test Errors)

第3章 線形回帰
-Linear regression-

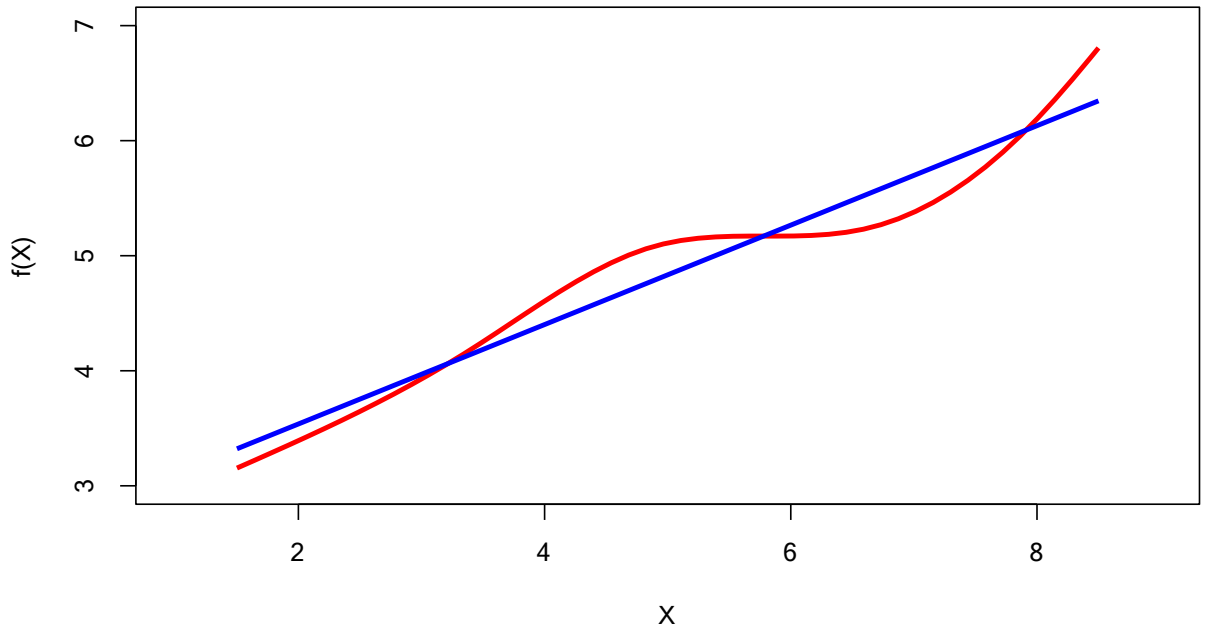
- 教師あり学習-supervised learning-
- 線形回帰-Linear regression-
- 単回帰
- 最小二乗(least squares)法
- 重線形回帰-Multiple Linear Regression-
- 前向き選択法-Forward selection-
- 後ろ向き選択法-Backward selection-
- モデル選択-Model selection-
- 質的予測変数-Qualitative Predictor-
- 交互作用-Interactions-
- 線形モデルの拡張

第3章 線形回帰
-Linear regression-

- 線形回帰は教師あり学習(supervised learning)への簡単なアプローチと云える. 線形モデルでは変数Yと変数群 $X_1, \mathrm{X}_2, \ldots, X_p$ との関係は線形となる.

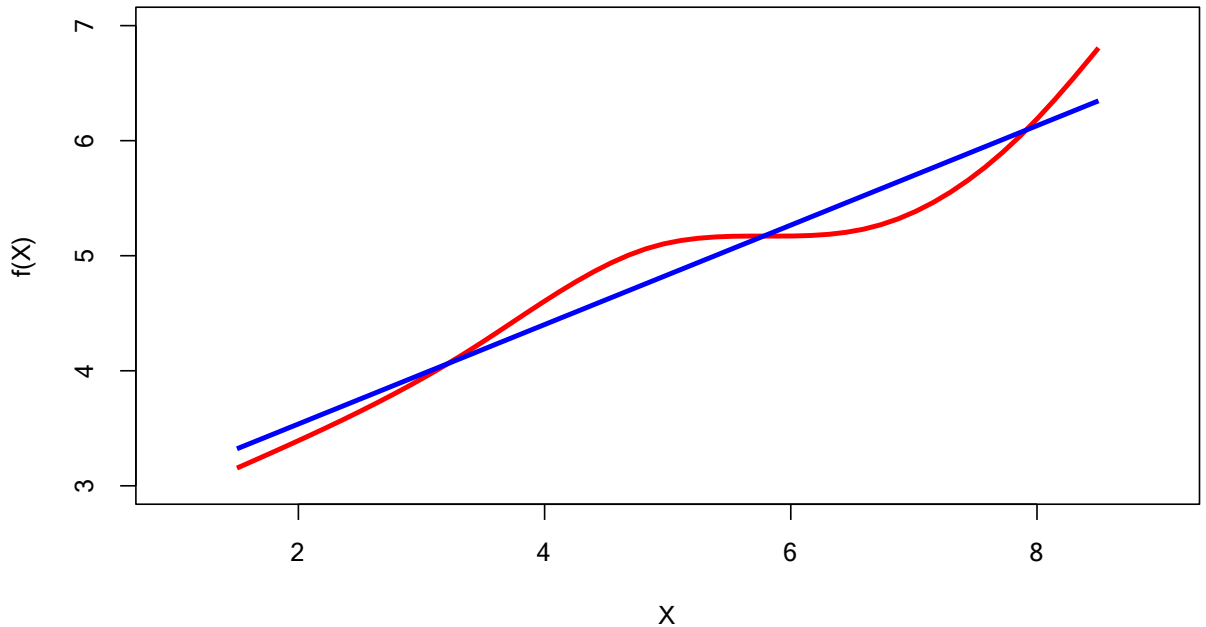
第3章 線形回帰
-Linear regression-

- 線形回帰は教師あり学習(supervised learning)への簡単なアプローチと云える. 線形モデルでは変数Yと変数群 $X_1, X_2, \ldots, X_p$ との関係は線形となる.
- しかし, 真の回帰関数は線形と云うことはないだろう!



第3章 線形回帰
-Linear regression-

- 線形回帰は教師あり学習(supervised learning)への簡単なアプローチと云える. 線形モデルでは変数Yと変数群 $X_1, \mathrm{X}_2, \ldots, X_p$ との関係は線形となる.
- しかし, 真の回帰関数は線形と云うことはないだろう!



- 線形と云う仮定は過度に楽観的過ぎると思える. しかし線形回帰は統計的概念としても実際的にも非常に有用である.

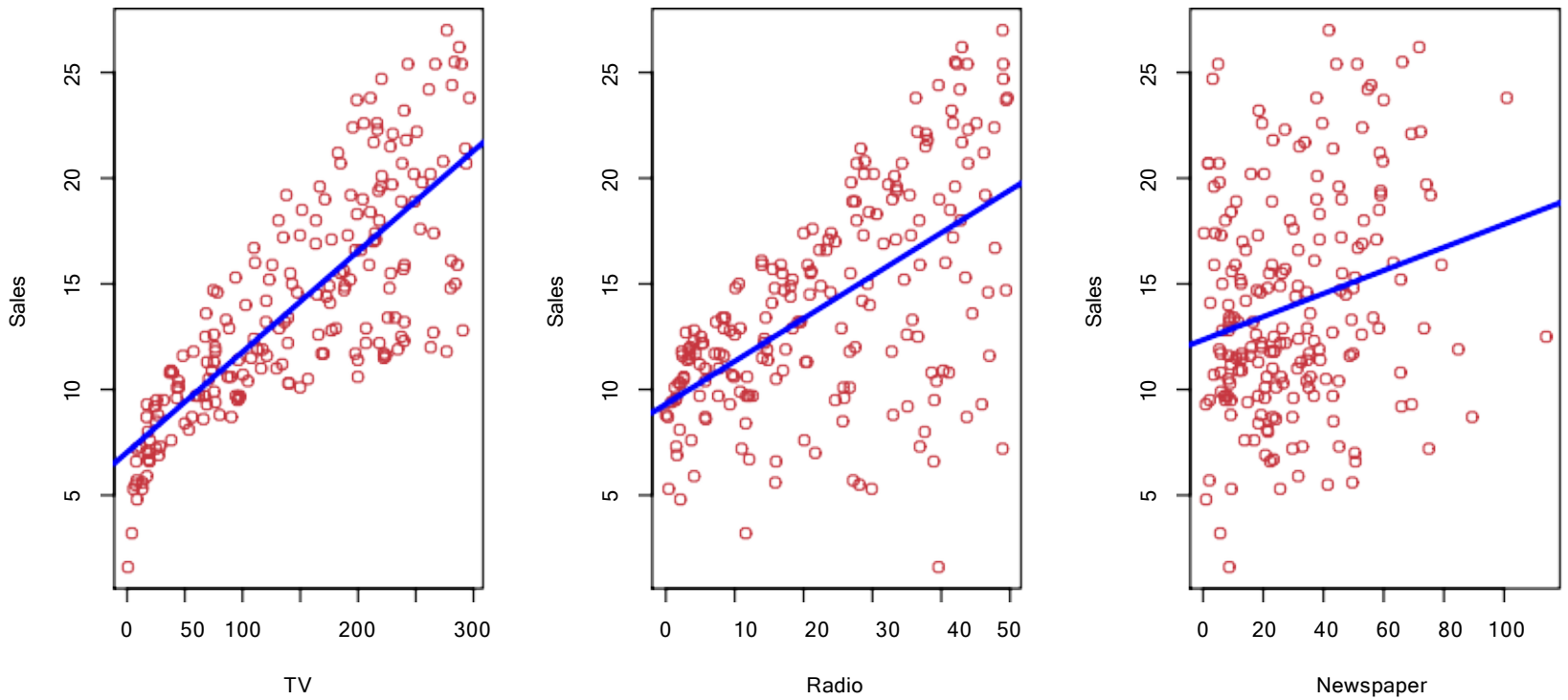
広告データへの線形回帰

次のスライドで示される広告データを考えよう.

ここで次のような疑問が生じる :

- 広告支出と販売量に関して関係があるだろうか?
- 広告支出と販売量の間の関係はどの程度強いだろうか?
- どのメディアによる公告が効果的だろうか?
- 将来の販売量をどの程度正確に予測できるだろうか?
- 関係は線形だろうか?
- 広告メディアの間に相乗効果(synergy)があるだろうか?

広告データ
(Advertising data)



単一の予測量Xを用いる単回帰 (Simple linear regression)

- 次の統計モデルを仮定する:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

ただしβ_0とβ_1は二つの未知の母数, 切片(*intercept*)と傾き(*slope*) ,は係数 (*coefficients*), あるいは母数(パラメター, *parameters*), ϵ は誤差項である.

- 統計モデルの係数 $\hat{\beta}_0, \hat{\beta}_1$ の推定値が与えられたとき次の式を用いて予測する

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

ここで $\hat{y}$ は変数 $X = x$とするときの変数 Y の予測値である. 記号のハット *hat* は推定値を意味する.

最小二乗(least squares)法による母数の推定

- ここで $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$を変数$X$の第$i$目の観測値に基づくの$Y$ の予測値としよう. このとき $e_i = y_i - \hat{y}_i$ は第 i番目の残差(*residual*)となる.

最小二乗(least squares)法による母数の推定

- ここで $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$を変数$X$の第$i$目の観測値に基づくの$Y$ の予測値としよう. このとき $e_i = y_i - \hat{y}_i$ は第 i番目の残差(*residual*)となる.
- 残差平方和(*residual sum of squares,* RSS) を定義すると

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

あるいは同じことであるが,

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

最小二乗(least squares)法による母数の推定

- ここで $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$を変数$X$の第$i$目の観測値に基づくの$Y$ の予測値としよう. このとき $e_i = y_i - \hat{y}_i$ は第 i番目の残差(*residual*)となる.
- 残差平方和(*residual sum of squares,* RSS) を定義すると

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

あるいは同じことであるが,

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- 最小二乗法は $\hat{\beta}_0$ と $\hat{\beta}_1$ を RSS を最小化するように選ぶ.

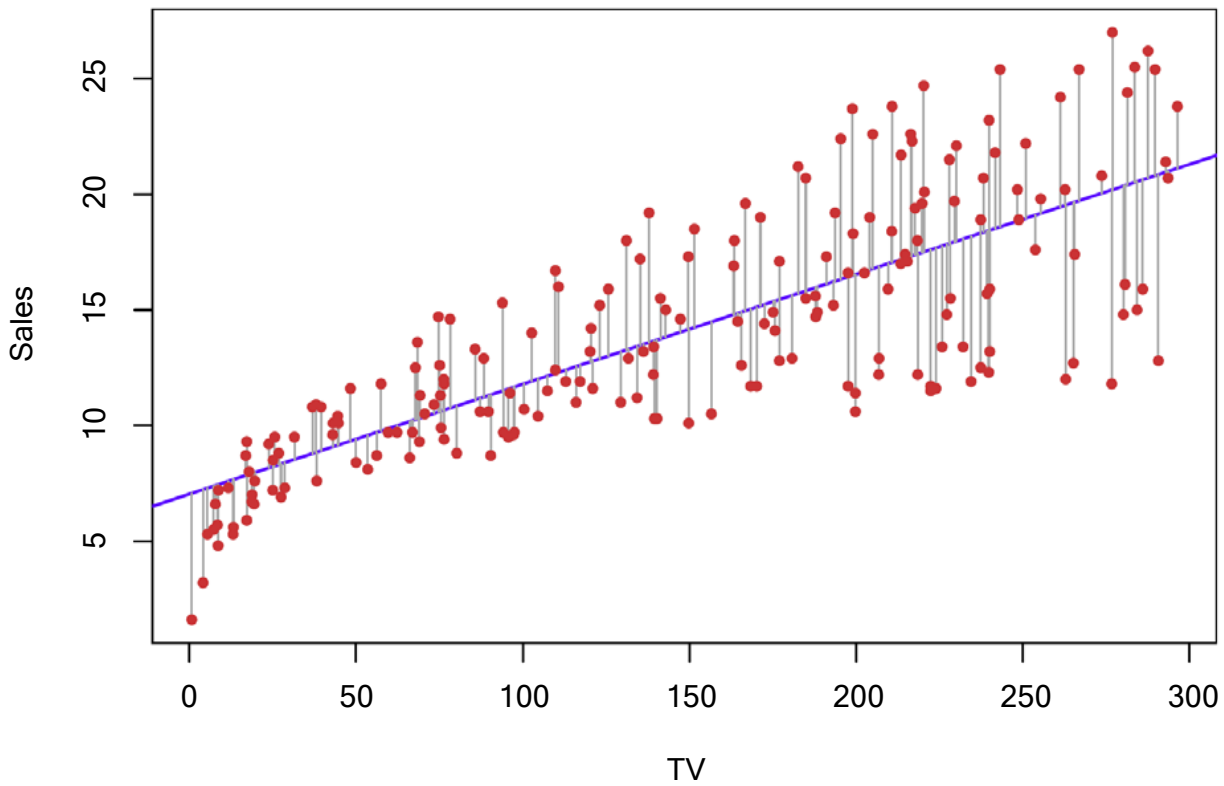
最小化する値は次のようになる

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

ただし $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^n y_i$, $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^n x_i$ は標本平均.

例: 広告データ



販売変数salesをテレビ変数TVへ最小二乗法で回帰. この場合には線形フィットにより変数間の重要な関係を把握しているが, プロットの左端の方には問題がある.

係数推定量の精度評価

- 推定量の標準誤差(standard error)は繰り返しサンプリングした時の変動を反映している. 次のように評価される

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

ただし$\sigma^2 = \mathrm{Var}(\epsilon)$

係数推定量の精度評価

- 推定量の標準誤差(standard error)は繰り返しサンプリングした時の変動を反映している. 次のように評価される

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

ただし$\sigma^2 = \mathrm{Var}(\epsilon)$

- この標本誤差を用いると信頼区間(*confidence intervals*)が計算できる. 例えば95% 信頼区間は確率 95%で真の未知母数を含む区間として求められる.
形式的には次の形になる:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

信頼区間(Confidence intervals): 続き

したがって, 近似的にチャンス 95% で区間

$$[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)]$$

は真値 β_1 (この例のような繰り返しサンプリングにより得られるという状況を仮定)を含む

信頼区間(Confidence intervals): 続き

したがって, 近似的にチャンス 95% で区間

$$[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)]$$

は真値 β_1 (この例のような繰り返しサンプリングにより得られるという状況を仮定)を含む

広告データの例では 95% 信頼区間(β_1)は [0.042, 0.053] となる.

仮説検定(Hypothesis testing)

- 標準誤差により係数についての仮説検定(*hypothesis tests*) を行うことができる. もっともよく考えられる仮説検定は次の帰無仮説(*null hypothesis*)

H_0 : 変数XとYの間には(線形)関係がない

versus.(対) 対立仮説(*alternative hypothesis)*

H_A : 変数XとYの関係がある.

仮説検定(Hypothesis testing)

- 標準誤差により係数についての仮説検定(*hypothesis tests*) を行うことができる. もっともよく考えられる仮説検定は次の帰無仮説(*null hypothesis*)

  H_0 : 変数XとYの間には(線形)関係がない

  versus.(対) 対立仮説(*alternative hypothesis)*

  H_A : 変数XとYの関係がある.

- 数理的には, この問題は仮説検定

$$H_0 : \beta_1 = 0$$

対(versus)

$$H_A : \beta_1 \neq 0,$$

となる. これはもし $\beta_1 = 0$ なら統計モデルは $Y = \beta_0 + \epsilon$ となり, 変数XはYと関係しないからである.

仮説検定(Hypothesis testing) ―続き

- 帰無仮説を検定するにはt統計量(*t-statistic*)を求める

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- この統計量は仮説$\beta_1 = 0$のもとで自由度$n - 2$のt分布(t-distribution)にしたがう.
- 統計ソフトを利用すると観測される $|t|$ に等しいかより大きくなる確率を簡単に計算できる. この確率をp値(*p-value*)と呼ぶ.

広告データによる結果

	係数	s.d.	t-値	p-値
切片 (Intercept)	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

統計モデルの全体の精度評価

- 残差標準誤差(*Residual Standard Error*)を求める:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

ただし残差平方和(*residual sum-of-squares*)は $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

統計モデルの全体の精度評価

- 残差標準誤差(*Residual Standard Error*)を求める:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

ただし残差平方和(*residual sum-of-squares*)は $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- *R二乗(R-squared),* あるいは説明された分散部分

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

ただし$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$は全変動和(*total sum of squares*).

統計モデルの全体の精度評価

- 残差標準誤差(*Residual Standard Error*)を求める:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

ただし残差平方和(*residual sum-of-squares*)は $\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- *R二乗(R-squared),* あるいは説明された分散部分

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

ただし$\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$は全変動和(*total sum of squares*).

- 単線形回帰モデルではrを変数Xと変数Yの相関係数とすると $R^2 = r^2$ となることを示せる:

$$r = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

広告データの結果

統計量	値
残差標準誤差	3.26
R^2	0.612
F-値statistic	312.1

重線形回帰(Multiple Linear Regression)

- 次の統計モデルを考える:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- ここで切片β_jは(すべての他の変数を固定したとき)変数X_jが1単位変化したもとで変数Yの平均的効果を示している. 広告モデルの例では次のようになる:

$$\text{広告(sales)} = \beta_0 + \beta_1 \times \text{テレビ(TV)} + \beta_2 \times \text{ラジオ(radio)} + \beta_3 \times \text{新聞(newspaper)} + \epsilon$$

回帰係数の解釈

- 理想的シナリオは予測変数は互いに無相関となる場合
  — バランスしたデザイン(*balanced design*) :
  - 各係数をそれぞれ個別に推定・検定できる.
  - *“他の変数を固定して係数β_jの変数X_jの1単位の変化がYに及ぼす影響という解釈”が可能.*
- 予測変数間に相関があると幾つかの問題を生じうる:
  - すべての係数の分散は増加, ときどき非常に大きくない.
  - 解釈が困難になりうる — 変数X_jが変化するとその変数以外のあらゆることが変わりうる.
- 観測データ分析からの因果性の主張(*Claims of Causality*)は避けるべきだろう.

回帰係数を解釈する上の落とし穴(woes)

"Data Analysis and Regression" Mosteller and Tukey 1977

- 回帰係数(regression coefficient) β_j は他の予測変数を固定したとき(*with all other predictors held fixed*)とき, 変数 X_jの1単位変化あたりに期待される変数 Y の変化を示している. しかし実際にはともに変化することが多い!

回帰係数を解釈する上の落とし穴(woes)

"Data Analysis and Regression" Mosteller and Tukey 1977

- 回帰係数(regression coefficient) β_j は他の予測変数を固定したとき(*with all other predictors held fixed*)とき, 変数 X_jの1単位変化あたりに期待される変数 Y の変化を示している. しかし実際にはともに変化することが多い!
- 例: Y ポケットの中の小銭の総額;
  X_1 = コインの数; X_2 = 1セント硬貨(pennies)数, 5セント硬貨(nickels)数,10セント硬貨(dimes)数. 変数Yの変数X_2への回帰係数は正, では統計モデルにおいてX_1の係数は？

回帰係数を解釈する上の落とし穴(woes)

"Data Analysis and Regression" Mosteller and Tukey 1977

- 回帰係数(regression coefficient) β_j は他の予測変数を固定したとき(*with all other predictors held fixed*)とき, 変数 X_jの1単位変化あたりに期待される変数 Y の変化を示している. しかし実際にはともに変化することが多い!
- 例: Y ポケットの中の小銭の総額;
  X_1 = コインの数; X_2 = 1セント硬貨(pennies)数, 5セント硬貨(nickels)数,10セント硬貨(dimes)数. 変数Yの変数X_2への回帰係数は正, では統計モデルにおいてX_1の係数は ?
- Y = あるシーズンでのあるアメフト選手のタックル数; W と H は体重と身長とする. フィットした回帰モデルは $\hat{Y} = b_0 + .50W - .10H$. Yの係数 $\hat{\beta}_2 < 0$であったら解釈は?

著名な統計家からの引用

“本質的には全ての統計モデルは正しくない, しかし役に立つモデルはある(Essentially, all models are wrong, but some are useful)”

ジョージ・ボックス(George Box)

著名な統計家からの引用

“本質的には全ての統計モデルは正しくない, しかし役に立つモデルはある(Essentially, all models are wrong, but some are useful)”

ジョージ・ボックス(George Box)

“複雑なシステムに変動がある場合にシステムがどう変動するかを知るには消極的に観察するだけではなくシステムを変動させてみるしかない.
The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively”

フレデリック・モステラー/ジョン・チューキー(Fred Mosteller/John Tukey, ボックスを引用して)

多重回帰の推定と予測

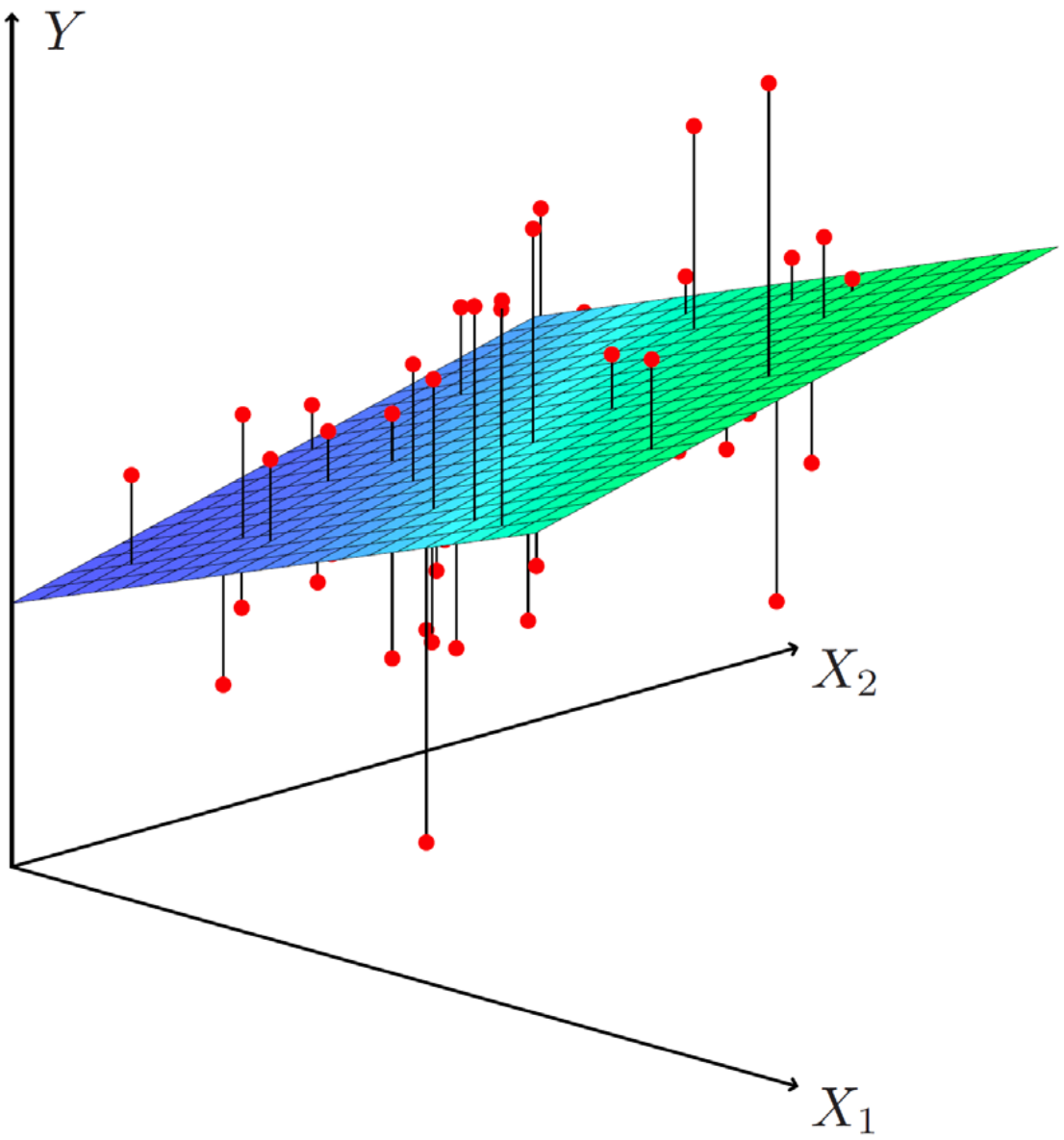
- 係数推定値$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$が得られた時, 次の公式を使って予測できる

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

- 係数 $\beta_0, \beta_1, \ldots, \beta_p$ は残差二乗和を最小にする:

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2 \end{aligned}$$

この計算は標準的ソフトで可能である. RSS を最小化する推定値 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$は最小二乗推定値と呼ばれる.


Y
X_2
X_1

広告データからの結果

	係数	Std. Error	t-統計量	p-値
切片	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

相関係数:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

幾つかの重要な疑問

1. *予測変数 $X_1, X_2, \ldots, X_p$ の中で少なくとも一つは目的変数にお予測に役立つだろうか？*

幾つかの重要な疑問

1. *予測変数 $X_1, X_2, \ldots, X_p$ の中で少なくとも一つは目的変数にお予測に役立つだろうか？*

2. *予測変数のすべてが変数 Y を説明するのに役立つか，あるいは一部分の変数にも役立つだろうか？*

幾つかの重要な疑問

1. *予測変数 $X_1, X_2, \ldots, X_p$ の中で少なくとも一つは目的変数にお予測に役立つだろうか？*

2. *予測変数のすべてが変数 Y を説明するのに役立つか，あるいは一部分の変数にも役立つだろうか？*

3. *どの程度統計モデルはデータにフィットするか？*

幾つかの重要な疑問

1. *予測変数 $X_1, X_2, \ldots, X_p$ の中で少なくとも一つは目的変数にお予測に役立つだろうか？*

2. *予測変数のすべてが変数 Y を説明するのに役立つか，あるいは一部分の変数にも役立つだろうか？*

3. *どの程度統計モデルはデータにフィットするか？*

4. *予測変数の値に対して目的変数の予測値は何か、またその予測はどの程度正確だろうか?*

少なくとも一つの予測変数は有用か?

この疑問に対して, F-統計量を用いることができる

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

統計量Quantity	値
残差標準誤差 Residual Standard Error	1.69
R^2	0.897
F-統計量	570

重要変数の決定

- 最も直接的なアプローチはサブセット回帰(*all subsets*あるいは *best subsets*): あらゆる部分集合に対して最小二乗フィットを計算, 訓練誤差とモデル・サイズをバランスさせる何らかの基準により選択することが一般的.

重要変数の決定

- 最も直接的なアプローチはサブセット回帰(*all subsets*あるいは *best subsets*): あらゆる部分集合に対して最小二乗フィットを計算, 訓練誤差とモデル・サイズをバランスさせる何らかの基準により選択することが一般的.

- しかしながら, しばしば可能なすべての統計モデルを調べることができないことがある, 例えば2^pの可能性がある; 例えば
$p = 40$のとき1億以上の可能性がある!
そこで部分集合の中で自動的な探索するアプローチが必要となる. そこで次に二つの良く利用されるアプローチを議論する.

前向き選択法(Forward selection)

- 帰無モデル(*null model*)より始める— 切片のみで予測変数がない統計モデル.
- p個の単線形回帰をフィットして帰無モデルにRSSが最小になる変数を加える.
- 次に全ての２変数モデルの中でRSSが最小になる変数を統計モデルに加える.
- この作業を何らかの停止ルールを満足するまで続ける. 例えば残ったすべての変数のp-値がある閾値以上となる etc.

後ろ向き選択法(Backward selection)

- 全ての予測変数を含む統計モデルから始める.
- P-値が最大となる変数を除く — すなわち統計的に最も有意とならない変数を除く.
- 次に$(p-1)$-個の変数を含むモデルをフィット, p-値が最大のとなる変数を除く.
- ある停止ルールが満たされるまで続ける. 例えば残ったすべての変数があらかじめ設定した閾値により有意となるときに止める.

モデル選択(Model selection) — 続き

- 後には前向き選択, 後ろ向き選択により選ばれる統計モデルにの中で最適”optimal”なモデルを選択するための基準について議論する.

- こうした基準として *Mallow's* C_p, 赤池情報量規準 *Akaike information criterion (AIC)*, ベイズ情報量規準 *Bayesian information criterion (BIC)*, 自由度修正済み決定係数 *adjusted* R^2 , 交差検証 *Cross-validation (CV)* などがある.

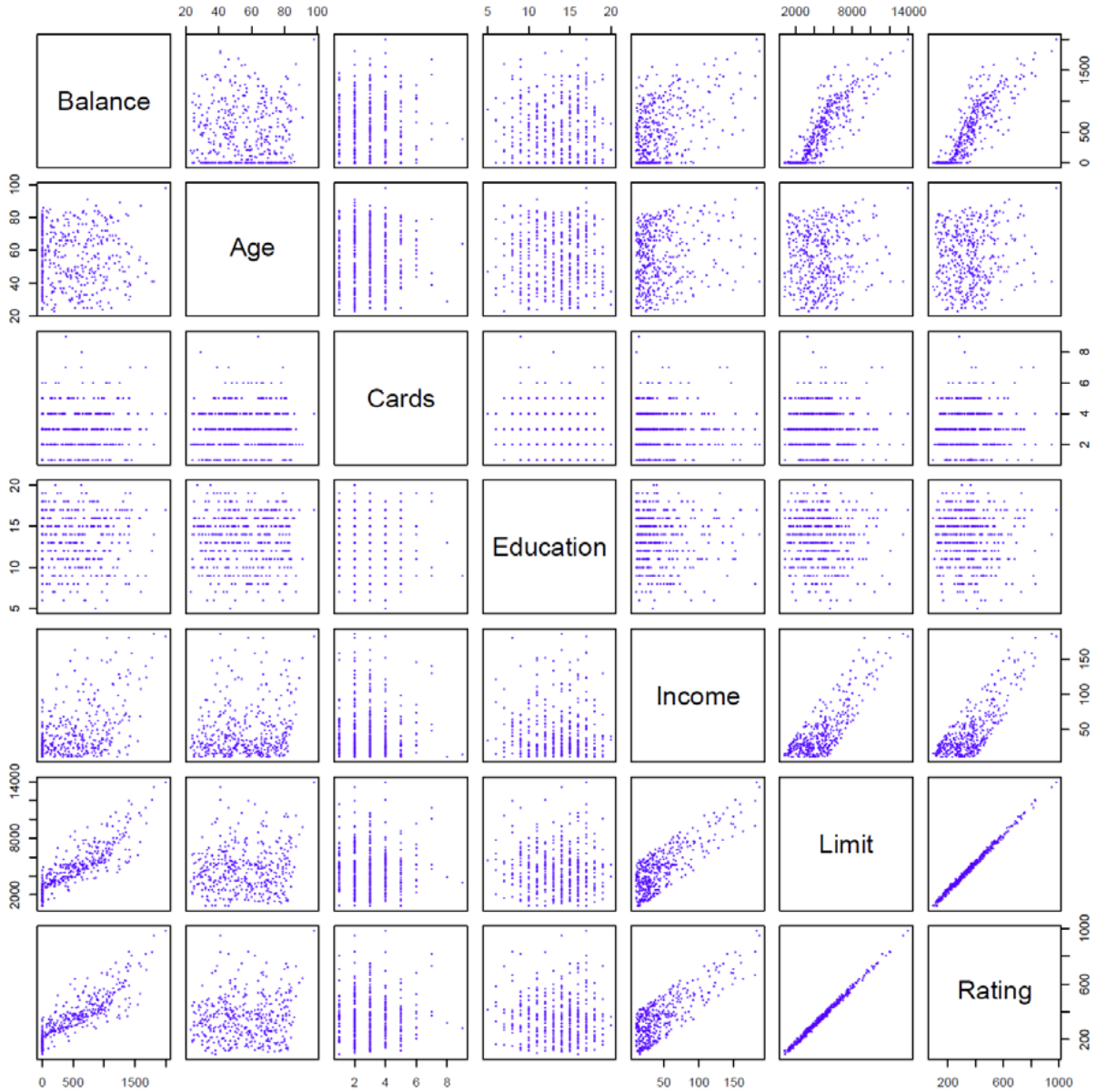
回帰モデルにおけるその他の問題

質的変数（Qualitative Predictors）

- 予測変数のいくつかは量的変数(*quantitative*)ではなく, 質的変数(*qualitative*)であり,幾つかの離散値をとることがある.
- これらの変数はカテゴリカル(*categorical*)予測変数, あるいはファクター変数(*factor variables*)と呼ぶ.
- 例えば次のスライドにあるクレジットカード・データの散布図を見ておこう.

7個の量的変数に加えて, 4個の質的変数がある: 性別(gender), 学生の身分(student), 婚姻状態(status), 人種(ethnicity)(白人(Caucasian), アフリカ系米国人(African American,AA) , アジア系(Asian)).

クレジットカード・データ



残高(Balance)
年齢(Age)
教育(Education)
所得(Income)
限界度(Limit)
評価(Rating)

質的予測変数(Qualitative Predictors)— 続き

例: クレジットカード・データにおいて, 他の変数を無視して男女間で残高の差を分析しよう. 新たな変数を導入する:

$$x_i = \begin{cases} 1 & \text{if } i\text{番目のデータは女性} \\ 0 & \text{if } i\text{番目のデータは男性} \end{cases}$$

結果として得られる統計モデル

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{番目のデータは女性} \\ \beta_0 + \epsilon_i & \text{if } i\text{番目のデータは男性} \end{cases}$$

モデルの解釈？

クレジットカード･データ ―続き

性別モデルの結果

	係数 Coefficient	Std. Error	t-統計量	p-値
切片	509.80	33.13	15.389	< 0.0001
性別[女性]	19.73	46.05	0.429	0.6690

2水準(level)以上の質的予測変数

- 2水準以上の質的変数の場合, 追加のダミー変数を作ればよい. 例えば民族性(ethnicity)変数については2個のダミー変数を作る. 最初の変数は

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{番目のデータはアジア系} \\ 0 & \text{if } i\text{番目のデータは非アジア系} \end{cases}$$

第2変数は

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{番目のデータは白人} \\ 0 & \text{if } i\text{番目のデータは非白人} \end{cases}$$

2水準(level)以上の質的予測変数—続き.

- 次に2変数を回帰方程式に利用, 次の統計モデルが得られる:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \\
&= \begin{cases}
\beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{番目のデータがアジア系} \\
\beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{番目のデータが白人} \\
\beta_0 + \epsilon_i & \text{if } i\text{番目のデータが}AA
\end{cases}
\end{aligned}
$$

- 常に水準(level)数より1少ないダミー変数が作れる.
- ダミー変数を持たない水準 — この例ではアフリカ系米国人 — はベースライン(*baseline*)となる.

民族性(ethnicity)についての結果

	係数	Std. Error	t-統計量	p-値
切片	531.00	46.32	11.464	< 0.0001
民族[アジア系]	-18.69	65.02	-0.287	0.7740
民族[白人]	-12.50	56.68	-0.221	0.8260

線形モデルの拡張

追加の仮定：交互作用(*interactions*)と非線形性(*nonlinearity*)の再考

交互作用(Interactions):

- これまでの広告データの分析では一つの広告メディアが売り上げを増加する効果は他のメディアにかける費用とは独立であることを仮定していた.
- 例えば, 線形モデル

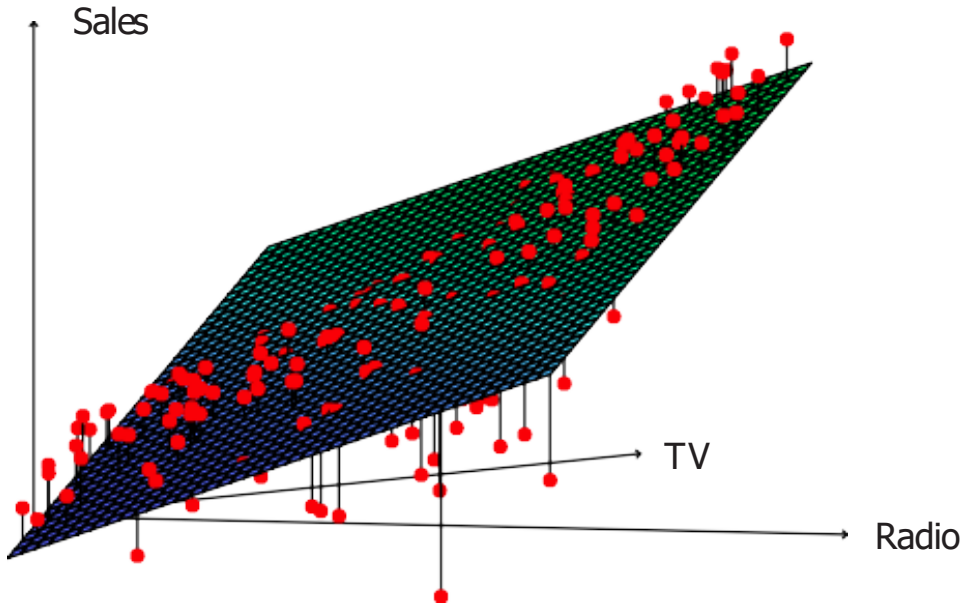
$$\widehat{\text{sales}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

はTV広告1単位の増加の売り上げへの平均効果はradioへの支出にいかんにかかわらず常にβ_1である.

交互作用(Interactions)—続き

- ここで仮にラジオ広告への支出増がTV広告の効果を増加させ, TV係数はラジオ支出増とともに増加するとしよう.
- この場合, 固定した予算$100, 000の下で半分をradio,半分をTVへの支出は全額をTV, あるいはラジオに支出するよりも売り上げ増加の効果が大きくなる可能性がある.
- マーケティング分野ではこの効果は相乗効果(シナジー, *synergy*)効果, 統計学では交互作用(*interaction*)と呼ばれている.

広告データにおける交互作用？



TV, あるいはラジオの水準が低ければ, 真の販売高は線形モデルによる予測よりも低い.

しかし広告を二つのメディアに分割すると線形モデルによる販売額を過少推定する傾向がある.

交互効果のモデル — 広告データ

統計モデルは次の形になる:

$$\begin{aligned}\text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times + \beta_3 \times (\text{radio} \times \text{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon.\end{aligned}$$

結果	係数 Coefficient	Std. Error	t-統計量	p-値
切片	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

解釈

- 表の結果は交互効果が重要であることを示唆している.
- 交互効果TV × radio のp-値はきわめて小さい, このことは仮説 $H_A: \beta_3 \neq 0$ を示唆している.
- 交互作用モデルの R^2 は 96.8%, 交互作用を含まないテレビとラジオによる販売額(sales)の予測するモデル 89.7% よりも良い.

解釈 — 続き

- このことは販売額の変動の (96.8−89.7)/(100−89.7) = 69% が統計モデルが交互作用のフィットにより説明されている.
- 表にある推定値はTV広告を$1, 000 増加すると販売額の増加は
  $\left(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}\right) \times 1000 = 19 + 1.1 \times \text{radio(1000)}$
- ラジオ広告 $1, 000の増加は販売額の増加
  $(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV(1000)}$
  に対応する.

階層-Hierarchy-

- 交互作用p-値は非常に小さいが. 関連する主効果(main effects,この場合はTVとラジオ)のp-値はそうでもないことがある.
- 階層原理(*hierarchy principle*) :

  もし交互作用項をモデルに含む場合には例えば係数の主効果が統計的に有意でなくても含めるべきである.

階層(Hierarchy)— 続き

- この原理の意味は交互作用項は主効果なしには解釈が困難(意味が変化する)だからである.
- 特に統計モデルが主効果がなければ, 交互作用項が主効果を含むことになる.

質的変数と量的変数の相互作用

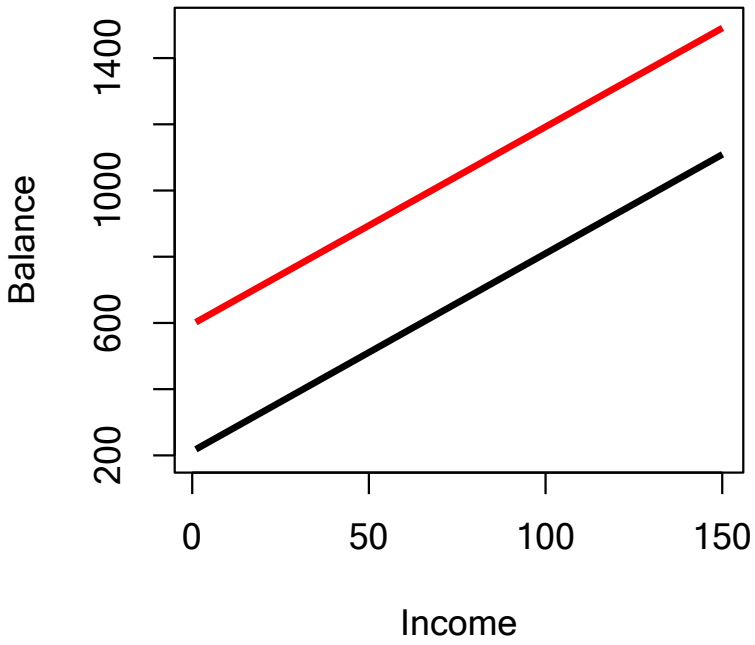
信用データ(Credit data)を利用して 所得(income,量的変数), 学生(student,質的変数)により残高(balance)を予測したいとする.

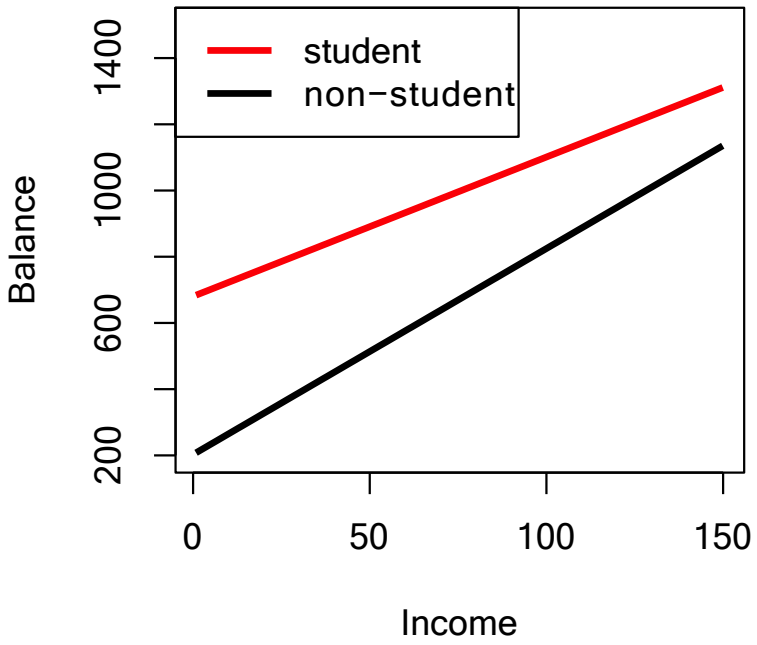
交互作用項がなければ統計モデルは次のようになる:

$$\begin{aligned}\text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases}\beta_2 & \text{if } i\text{番目は学生} \\ 0 & \text{if } i\text{番目は学生でない}\end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases}\beta_0 + \beta_2 & \text{if } i\text{番目は学生} \\ \beta_0 & \text{if } i\text{番目は学生でない}\end{cases}\end{aligned}$$

交互作用が存在すると次のようになる

$$\begin{aligned}
&\text{balance}_i \\
&\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if } i\text{番目は学生} \\ 0 & \text{if } i\text{番目は学生でない} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if } i\text{番目は学生} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if } i\text{番目は学生でない} \end{cases}
\end{aligned}$$

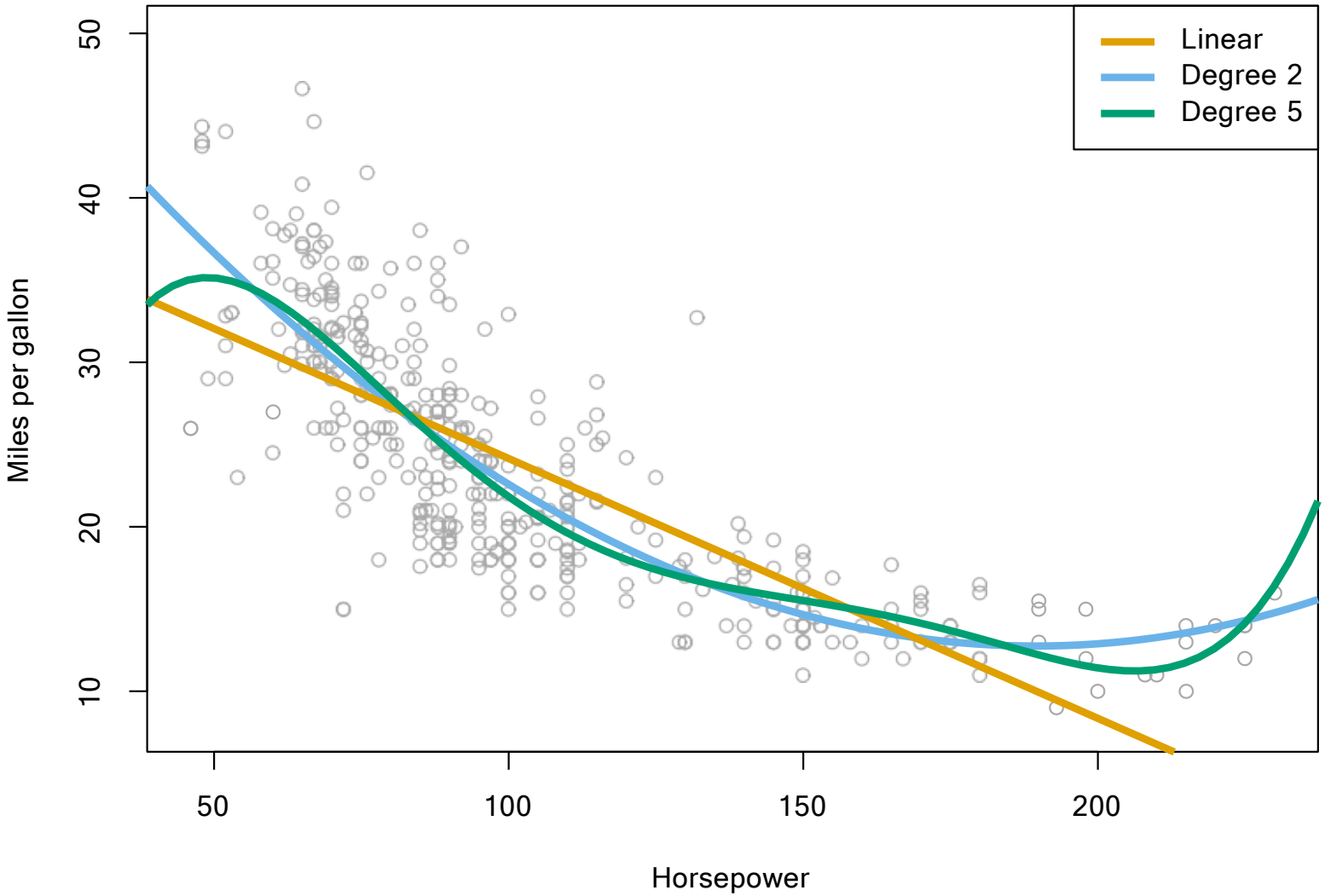




クレジット・データ; 左: 所得income)と学生(student)の交互作用なし. 右: 所得と学生身分の相互作用がある場合.

予測変数の非線形効果

Auto データへの多項式回帰



図によると

$$\text{mpg} = \beta_0 + \beta_1 \times \text{馬力(horsepower)} + \beta_2 \times \text{馬力(horsepower)}^2 + \epsilon$$

の方がフィットが良いかもしれない

	Coefficient	Std. Error	t-statistic	p-value
切片	56.9001	1.8004	31.6	< 0.0001
馬力 (horsepower)	-0.4662	0.0311	-15.0	< 0.0001
馬力^2 (horsepower)^2	0.0012	0.0001	10.1	< 0.0001

これまでに議論しなかった重要な内容

外れ値(Outliers)
誤差項の分散不均一性
高レバレッジ点(High leverage points)
共線性(Collinearity)

これらについては教科書3.3.3節を見よ

線形モデルの一般化

このコースでは線形モデルの利用を拡張した統計的方法をどのように適用するか議論していく:

- *分類問題(Classification problems):* ロジッティック回帰(logistic regression), サポート・ベクトル分類器(support vector machines)
- *非線形性(Non-linearity):* カーネル平滑化(kernel smoothing), スプライン(splines), 一般化加法モデル(generalized additive models); 最近傍法(nearest neighbor methods).
- *交互作用(Interactions):* 木モデル(Tree-based methods), バギング(bagging), ランダム・フォレスト(random forests), ブースティング(boosting) ,これらは非線形性も表現できる.
- *正則化フィット(Regularized fitting):* リッジ回帰(Ridge regression), ラッソ(lasso)

第4章　分類
−Classification−

- 質的変数
- 分類
- ロジスティック回帰
- 最尤法
- 判別分析
- 線形判別・二次判別
- ナイーブベイズ

第4章　分類
−Classification−

- 質的変数が

$$eye\ color \in \{brown, blue, green\}$$

$$email \in \{spam, ham\}$$

という順不同の集合Cから値を取る。
- 特徴ベクトルXと質的な応答変数Yが与えられたものとする。ただし、Yは集合Cから値を取ったものである。この分類問題の目的はXを入力とし、Yの値を予想する関数$C(X)$を作ることである。i.e., $C(X) \in C$.
- しばしば我々はXがC内のそれぞれのカテゴリに所属する確率を推定することにより興味を持つ

第4章　分類
−Classification−

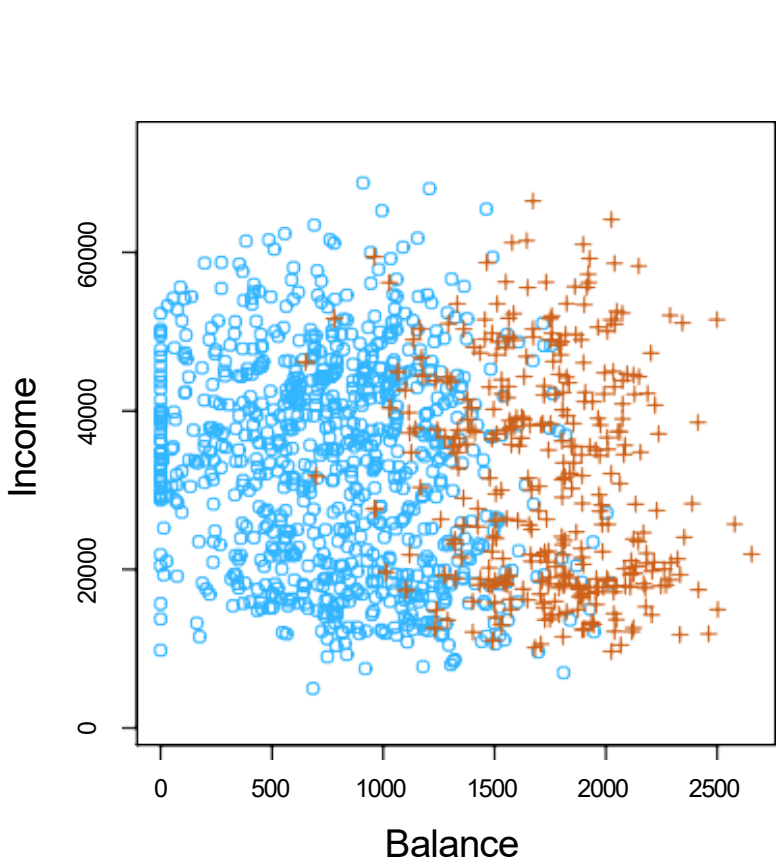
- 質的変数が

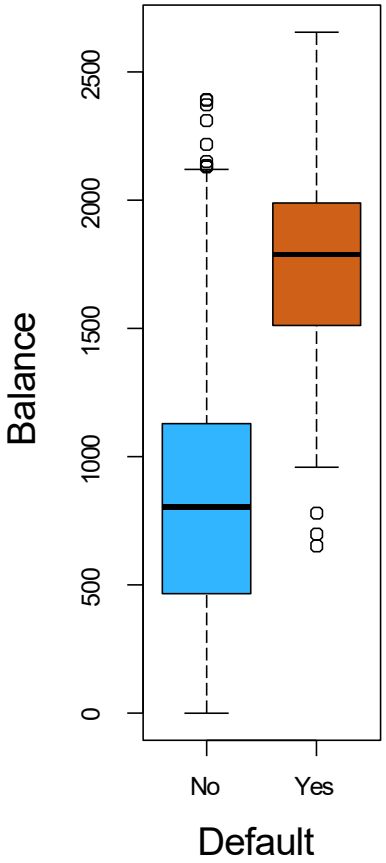
$$eye\ color \in \{brown, blue, green\}$$

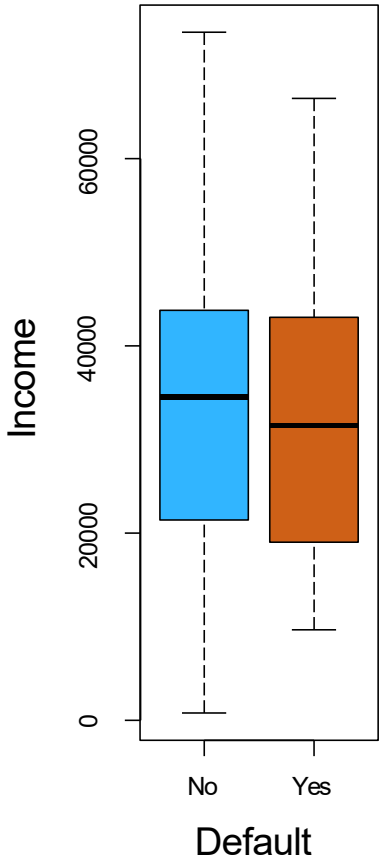
$$email \in \{spam, ham\}$$

という順不同の集合Cから値を取る。
- 特徴ベクトルXと質的な応答変数Yが与えられたものとする。ただし、Yは集合Cから値を取ったものである。この分類問題の目的はXを入力とし、Yの値を予想する関数$C(X)$を作ることである。i.e., $C(X) \in C$.
- しばしば我々はXがC内のそれぞれのカテゴリに所属する確率を推定することにより興味を持つ。
例えば保険のクレームが欺瞞的である可能性を推定することは、欺瞞的がそうでないかを分類することよりもより価値がある

例：クレジットカードの債務不履行データ







線形回帰が利用できるのか

仮にデフォルトの分類タスク

$$Y = \begin{cases} 0 \ \ if \ No \\ 1 \ \ if \ Yes \end{cases}$$

我々は単純にXに対するYの線形回帰を行うことができるか？
また、$\hat{y} > 0.5$の時、Yesと分類できるか？

線形回帰が利用できるのか

仮にデフォルトの分類タスク

$$Y = \begin{cases} 0 \ \ if\ No \\ 1 \ \ if\ Yes \end{cases}$$

我々は単純にXに対するYの線形回帰を行うことができるか？
また、$\hat{y} > 0.5$の時、Yesと分類できるか？

- 応答変数が2値変数の場合、線型回帰は分類器としてよく機能でき、のちに議論する線型判別分析と等しい。
- $E(Y|X = x) = \Pr(Y = 1 \mid X = x)$という母集団において、この分類の問題には回帰が最適と考えるかもしれない。

線形回帰が利用できるのか

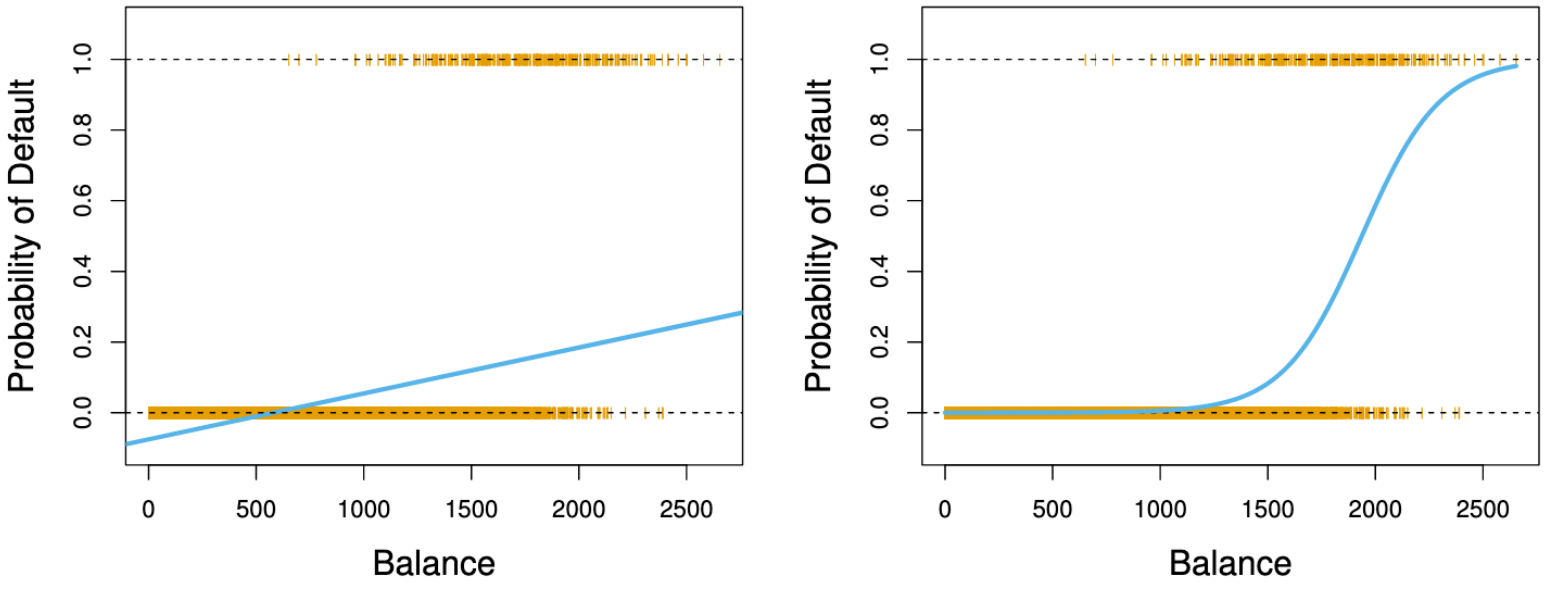
仮にデフォルトの分類タスク

$$Y = \begin{cases} 0 \ \ if\ No \\ 1 \ \ if\ Yes \end{cases}$$

我々は単純にXに対するYの線形回帰を行うことができるか？
また、$\hat{y} > 0.5$の時、Yesと分類できるか？

- 応答変数が2値変数の場合、線型回帰は分類器としてよく機能でき、のちに議論する線型判別分析と等しい。
- $E(Y|X = x) = \Pr(Y = 1 \mid X = x)$という母集団において、この分類の問題には回帰が最適と考えるかもしれない。
- しかしながら、線型回帰は0以下か1以上の確率を生成するかもしれない。ロジスティック回帰の方が適切である。

線形回帰とロジスティック回帰



このオレンジマーカーは応答変数Yを0か1の値にしたものである。線型回帰は$\Pr(Y = 1 \mid X)$をうまく推定できない。ロジスティック回帰はこの問題にうまくあっているように見える。

線形回帰（つづき）

今我々は三つの可能な値を取る応答変数を持つと仮定する。緊急外来に患者がきて、我々は彼らを症状に応じて分類しなければならない。

$$Y = \begin{cases} 1 & \text{脳卒中の場合;} \\ 2 & \text{薬物過剰摂取の場合;} \\ 3 & \text{てんかん発作の場合.} \end{cases}$$

このコーディングは、脳卒中と薬物過剰摂取の違いが、薬物過剰摂取とてんかん発作の違いと同じであることを示唆するものである

線形回帰（つづき）

今我々は三つの可能な値を取る応答変数を持つと仮定する。緊急外来に患者がきて、我々は彼らを症状に応じて分類しなければならない。

$$Y = \begin{cases} 1 & \text{脳卒中の場合;} \\ 2 & \text{薬物過剰摂取の場合;} \\ 3 & \text{てんかん発作の場合.} \end{cases}$$

このコーディングは、脳卒中と薬物過剰摂取の違いが、薬物過剰摂取とてんかん発作の違いと同じであることを示唆するものである。

線形回帰はここでは適切ではない。

マルチクラスロジスティック回帰または判別分析がより適切である。

ロジスティック回帰

略して$p(X) = \Pr(Y = 1 \mid X)$と書いて、予測変数balanceを使って応答変数defaultを予測してみよう。

ロジステック回帰はこの形式を使う。

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

($e \approx 2.71828$は定数　オイラー数)

β_0, β_1, Xがどんな値であろうと、$p(X)$は0から1の間の値を取ることは容易にわかる。

ロジスティック回帰

略して$p(X) = \Pr(Y = 1 \mid X)$と書いて、予測変数balanceを使って応答変数defaultを予測してみよう。

ロジステック回帰はこの形式を使う。

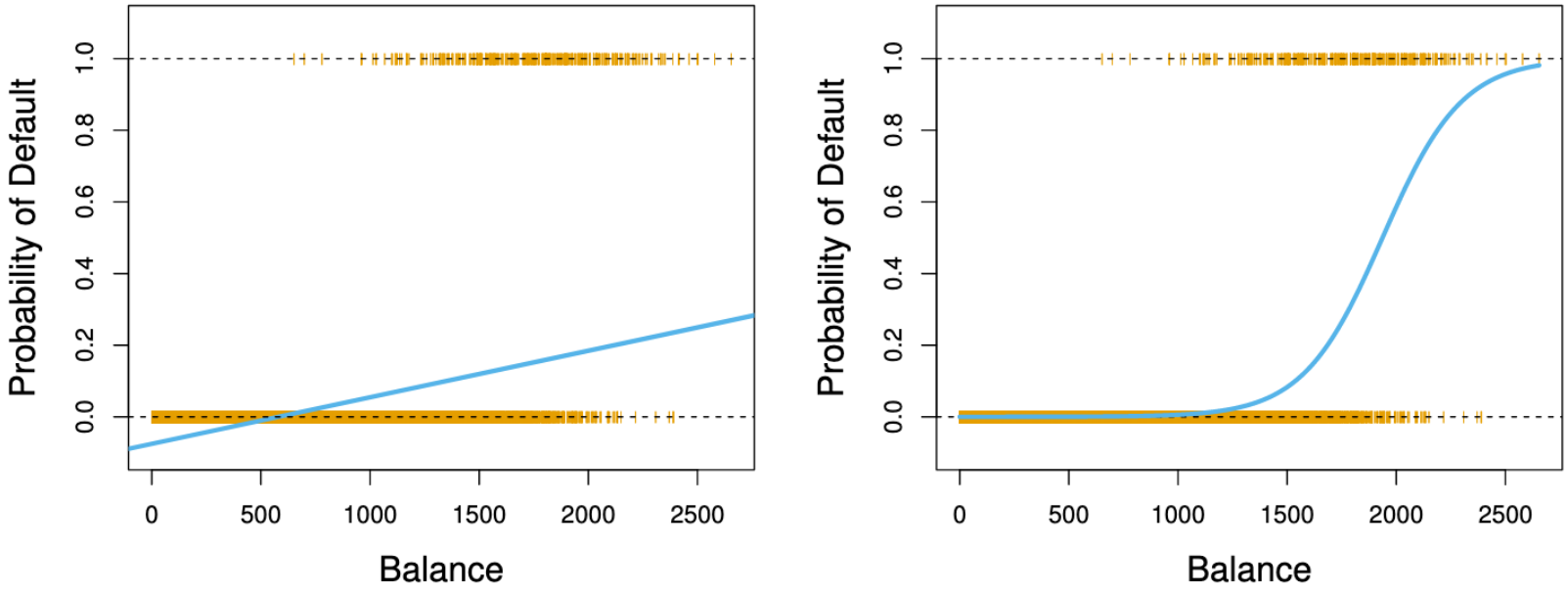
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

($e \approx 2.71828$は定数　オイラー数)
β_0, β_1, Xがどんな値であろうと、$p(X)$は0から1の間の値を取ることは容易にわかる。
上記の式を少し変形することにより

$$\log\left(\frac{P(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

を得る。この単調変換は$p(X)$のlog oddsまたはlogit変換と呼ばれる。(logは自然対数を意味する)

線形回帰とロジスティック回帰



ロジステック回帰では$p(X)$が0と1の間の値をとることが保証される。

最尤法

最尤法を用いてパラメータを推定する。

$$l(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

この尤度は、観測されたデータについて0と1の確率を与える。
観測されたデータの尤度が最大になるようにβ_0とβ_1を選ぶ。

最尤法

最尤法を用いてパラメータを推定する。

$$l(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$$

この尤度は、観測されたデータについて0と1の確率を与える。
観測されたデータの尤度が最大になるようにβ_0とβ_1を選ぶ。

ほとんどの統計パッケージは、最尤法による線形ロジスティック回帰モデルのあてはめが可能である。Rでは、glm関数を使用する。

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
balance	0.0055	0.0002	24.9	< 0.0001

予測

債務残高(balance)1000ドルの人がdefaultになる確率はどれくらいでしょうか?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+ e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+ e^{-10.6513+0.0055\times 1000}} = 0.006$$

予測

債務残高(balance)1000ドルの人がdefaultになる確率はどれくらいでしょうか?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+ e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+ e^{-10.6513+0.0055\times 1000}} = 0.006$$

2000ドルの残高の場合は？

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+ e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+ e^{-10.6513+0.0055\times 2000}} = 0.586$$

今度はstudentを予測変数に使って、もう一度やってみよう。

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-3.5041	0.0707	-49.55	< 0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

今度はstudentを予測変数に使って、もう一度やってみよう。

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-3.5041	0.0707	-49.55	< 0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

$$\widehat{Pr}(default = Yes|student = Yes) = \frac{e^{-3.5041+0.4049\times1}}{1 + e^{-3.5041+0.4049\times1}} = 0.0431$$

$$\widehat{Pr}(default = Yes|student = No) = \frac{e^{-3.5041+0.4049\times0}}{1 + e^{-3.5041+0.4049\times0}} = 0.0292$$

複数の変数を用いたロジスティック回帰

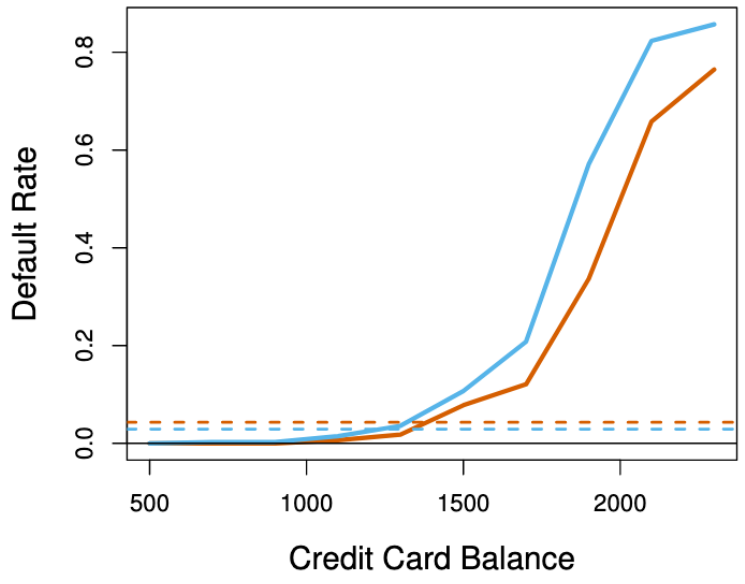
$$\log \frac{p(X)}{1-p(X)} = \beta_0 + \beta_1 X_1 + \cdots\cdots + \beta_p X_p$$

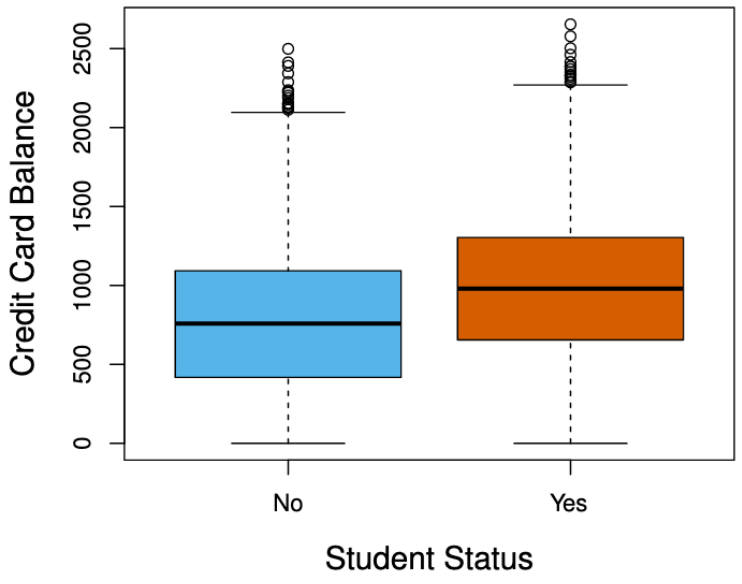
$$p(X) = \frac{e^{\beta_o + \beta_o X_1 + \cdots\cdots + \beta_p X_p}}{1 + e^{\beta_o + \beta_o X_1 + \cdots\cdots + \beta_p X_p}}$$

	Coefficient	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

学生の係数が、以前はプラスだったのに、なぜマイナスになったのですか？

交絡

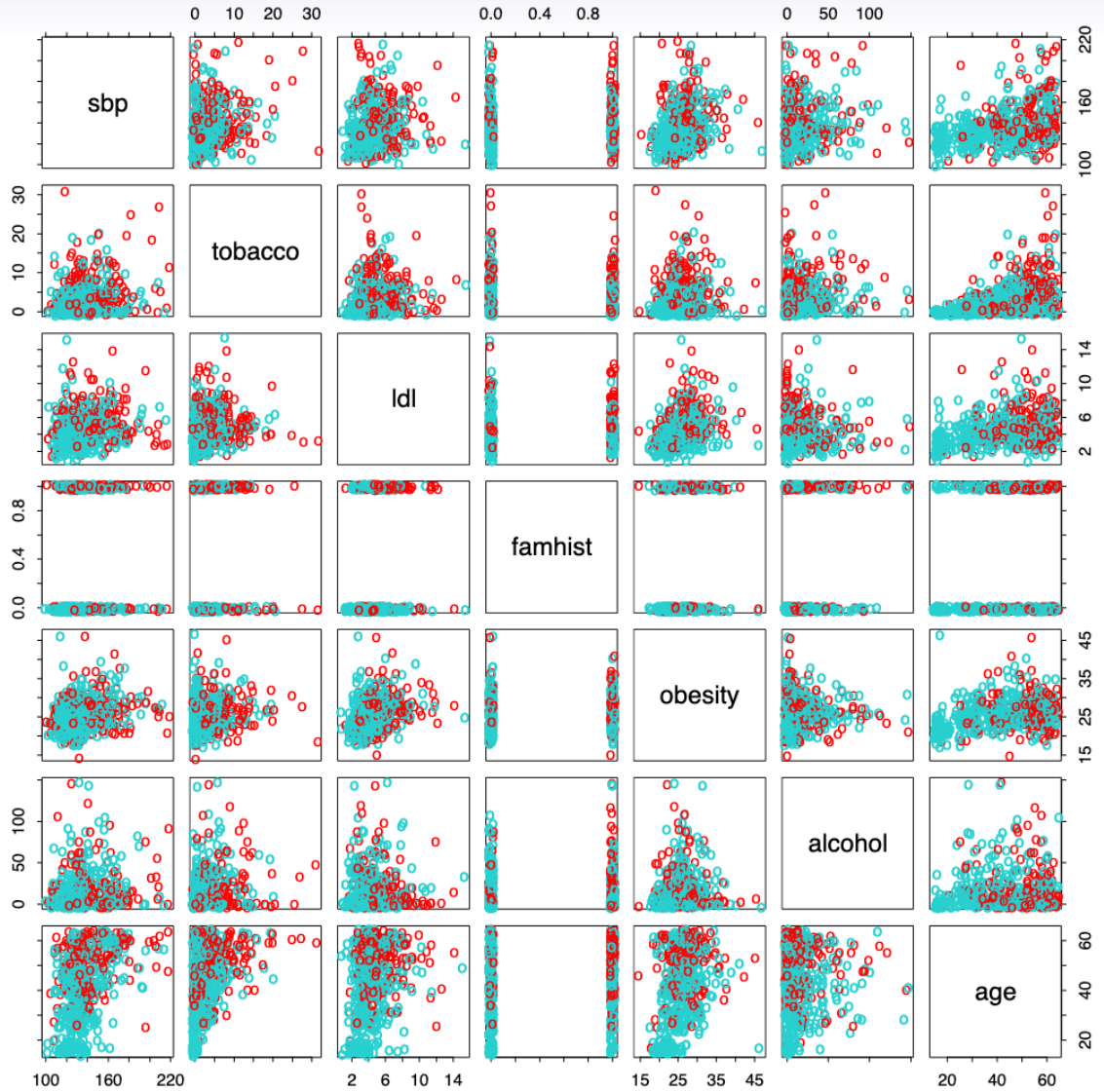




- 学生の顧客（オレンジ）は学生でない顧客（青）より債務残高が多い傾向があるので、平均的な債務不履行の確率は非学生より高くなる。
- しかし、債務残高のどのレベルにおいても、学生の顧客は学生でない顧客より債務不履行の確率が低い。

例: 南アフリカの心臓病データ

- 80年代前半に南アフリカの西ケープ州で発生した心筋梗塞160例と対照302例（すべて男性、年齢15-64歳）。
- この地域の全体的な有病率は非常に高い：5.1%。
- 7つの予測因子（危険因子）を測定し、散布図に示した。
- 目標は、危険因子の相対的な強さと方向を特定することである。
- これは、より健康的な食事について一般の人々を教育することを目的とした介入研究の一部であった。



南アフリカ心臓病データの散布図。応答変数は色分けされており、心筋梗塞は赤、対照は青緑である。famhistは2値変数で、1は心筋梗塞の家族歴を示す。

```
> heartfit<-glm(chd~.,data=heart,family=binomial)
> summary(heartfit)

Call:
glm(formula = chd ~ ., family = binomial, data = heart)

Coefficients:
                    Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)      -4.1295997   0.9641558   -4.283  1.84e-05 ***
Sbp収縮期血圧     0.0057607   0.0056326    1.023   0.30643
Tobaccoたばこ     0.0795256   0.0262150    3.034   0.00242 **
Ldlコレステロール  0.1847793   0.0574115    3.219   0.00129 **
famhistPresent    0.9391855   0.2248691    4.177  2.96e-05 ***
Obesity肥満      -0.0345434   0.0291053   -1.187   0.23529
Alcoholアルコール  0.0006065   0.0044550    0.136   0.89171
Age年齢           0.0425412   0.0101749    4.181  2.90e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11   on 461  degrees of freedom
Residual deviance: 483.17   on 454  degrees of freedom
AIC: 499.17
```

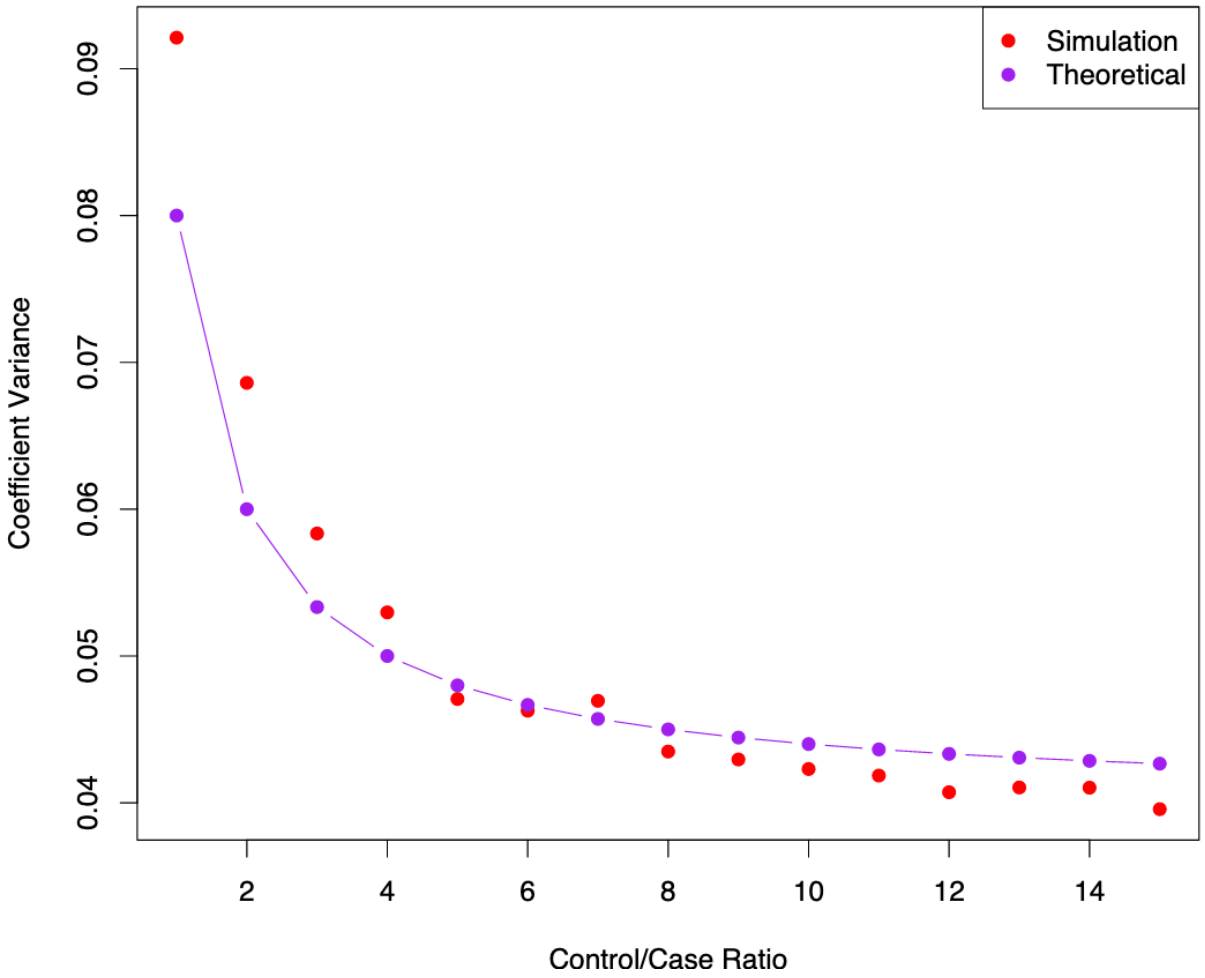
ケースコントロールサンプリングとロジスティック回帰

- 南アフリカのデータでは、160人の症例、302人の対照者--$\tilde{\pi}$＝0.35が症例である。しかし、この地域の心筋梗塞の有病率は、$\pi = 0.05$です。
- ケース・コントロール標本を用いると、回帰パラメータβ_jを正確に推定することができる（我々のモデルが正しい場合）；定数項 β_0は不正確である。
- 推定された切片は、簡単な変換によって修正できる。

$$\hat{\beta}_0^* = \hat{\beta}_0 + \log\frac{\pi}{1-\pi} - \log\frac{\tilde{\pi}}{1-\tilde{\pi}}$$

- 稀なケースも多いので、すべて受ける。その5倍までのコントロールで十分である。次のフレームを見る。

アンバランスな2値変数における収穫逓減



症例よりも対照を多くサンプリングすることで、パラメータ推定値の分散を減少させることができます。しかし、約5対1の比率になると分散の減少は平坦になる。

2つ以上のクラスを持つロジスティック回帰

これまで、ロジスティック回帰を2クラスで説明した。この手法は、2つ以上のクラスへ簡単に一般化できる。たとえば、Rのglmnetパッケージでは以下の式が用いられる。

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots\cdots+\beta_{pk}X_p}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l}X_1+\cdots\cdots+\beta_{pl}X_p}}$$

ここで、各クラスに一次関数が存在するとする。

2つ以上のクラスを持つロジスティック回帰

これまで、ロジスティック回帰を2クラスで説明した。この手法は、2つ以上のクラスへ簡単に一般化できる。たとえば、Rのglmnetパッケージでは以下の式が用いられる。

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots\cdots+\beta_{pk}X_p}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l}X_1+\cdots\cdots+\beta_{pl}X_p}}$$

ここで、各クラスに一次関数が存在するとする。
(2クラスのロジスティック回帰と同様に、K−1個の線形関数のみが必要であることに気づくでしょう。)

2つ以上のクラスを持つロジスティック回帰

これまで、ロジスティック回帰を2クラスで説明した。この手法は、2つ以上のクラスへ簡単に一般化できる。たとえば、Rのglmnetパッケージでは以下の式が用いられる。

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\cdots\cdots+\beta_{pk}X_p}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l}X_1+\cdots\cdots+\beta_{pl}X_p}}$$

ここで、各クラスに一次関数が存在するとする。
(2クラスのロジスティック回帰と同様に、K−1個の線形関数のみが必要であることに気づくでしょう。)
マルチクラスロジスティック回帰は、 多項ロジスティック回帰とも呼ばれる。

判別分析

ここでは、与えられた各応答変数Yについて、予測変数Xの確率分布をモデル化する。そしてベイズの定理で条件の対応を反転させ、$\Pr(Y|X)$を推定する。
これらの確率分布に正規分布を用いると、線形あるいは2次判別分析になる。
しかし、この方法は非常に一般的であり、他の分布を用いることも可能である。ここでは、正規分布を用いる。

分類のためのベイズの定理

ベイズは有名な数学者であり、その名前は統計的・確率的モデリングの大きなサブフィールドを代表するものである。ここでは、ベイズの定理として知られる簡単な結果に注目する。

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

分類のためのベイズの定理

ベイズは有名な数学者であり、その名前は統計的・確率的モデリングの大きなサブフィールドを代表するものである。ここでは、ベイズの定理として知られる簡単な結果に注目する。

$$\Pr(Y = k|X = x) = \frac{\Pr(X = x|Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

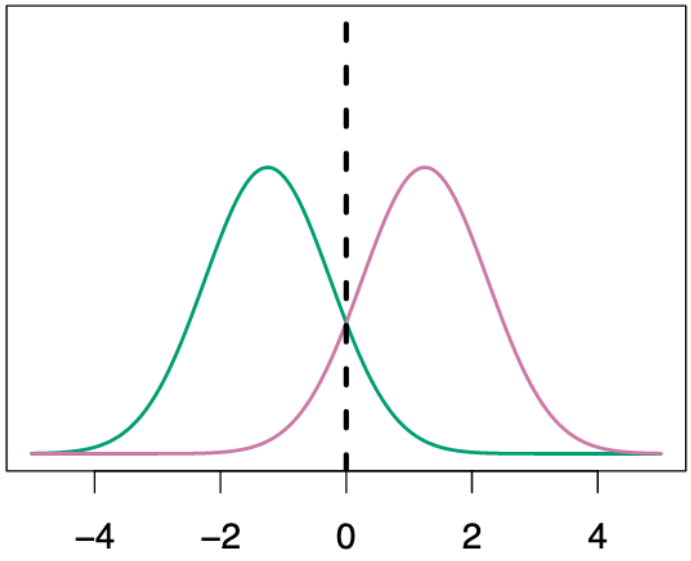
判別分析では少し違った書き方をする。

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

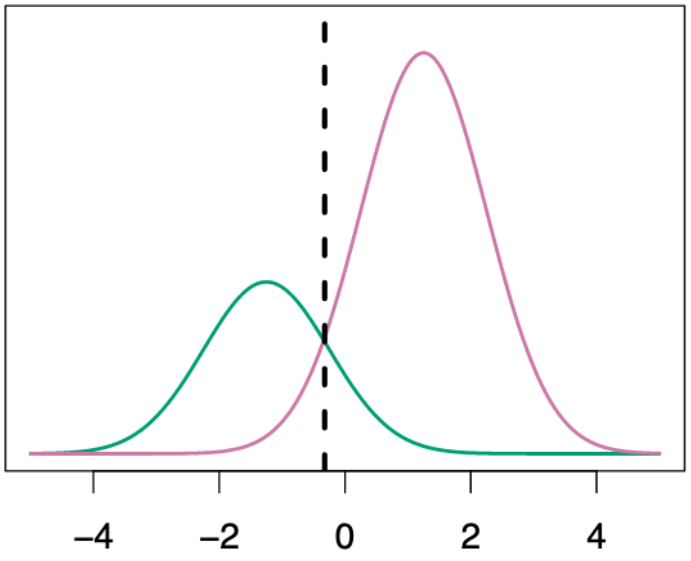
$f_k(x) = \Pr(X = x|Y = k)$は$k$番目のクラスから得られた観測値$X = x$の確率密度である。ここでは、これらの密度関数に関して正規性を仮定する。
$\pi_k = \Pr(Y = k)$はクラスkに属する周辺確率または事前確率である。

最も高い密度に応じて分類する

π_1=.5, π_2=.5

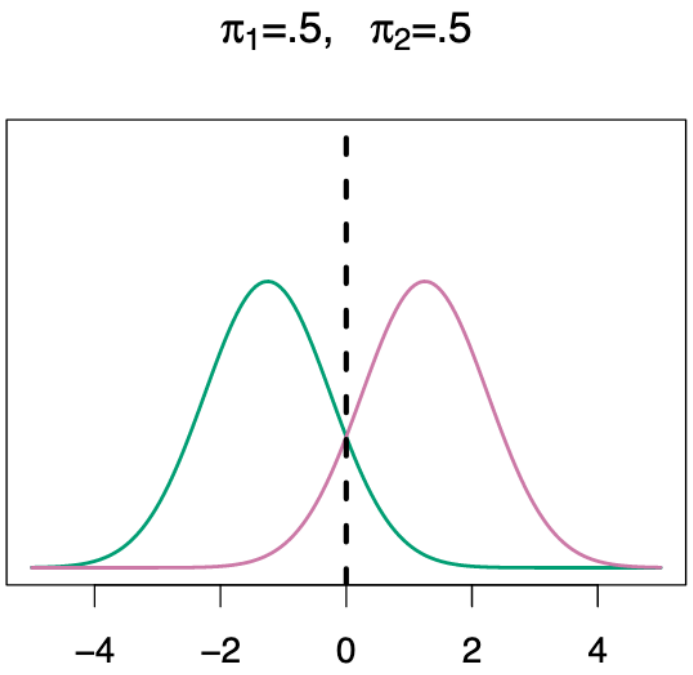


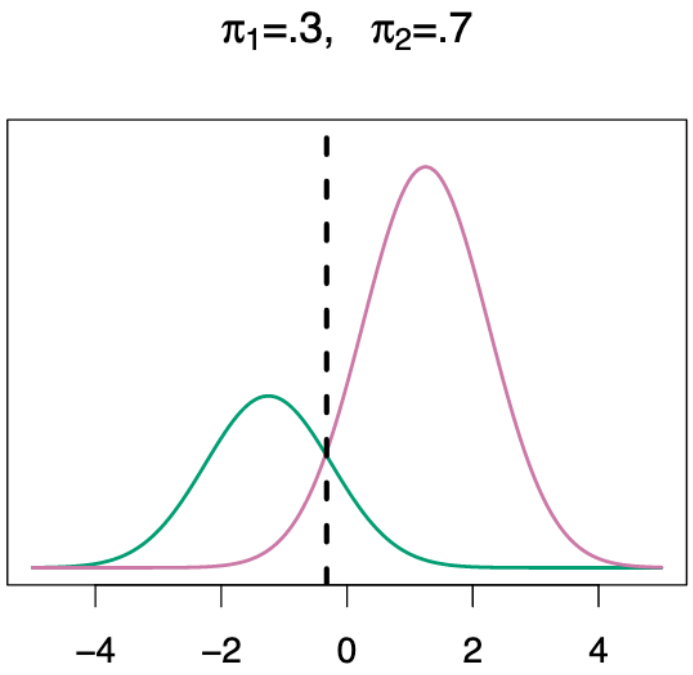
π_1=.3, π_2=.7



どの密度が最も高いかによって、新しい点を分類する。

最も高い密度に応じて分類する





どの密度が最も高いかによって、新しい点を分類する。
事前確率密度関数が異なる場合は、それも考慮して$\pi_k f_k(x)$を比較する。右側では、ピンクのクラスが有利で、判定境界が左に移動している。

なぜ判別分析なのか？

- クラスがよく分離されている場合、ロジスティック回帰モデルのパラメータ推定値は驚くほど不安定になる。線形判別分析はこの問題に悩まされることはない。
- nが小さく、予測変数Xの分布が各クラスでほぼ正規分布であれば、やはりロジスティック回帰モデルより線形判別モデルの方が安定である。
- 線形判別分析は、データを低次元で見ることもできるため、2つ以上のクラスがある場合によく利用される。

$p = 1$の場合の線形判別分析

正規分布の確率密度関数:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

ここで、μ_kと$\sigma_k{}^2$はそれぞれk番目のクラスの平均と分散である。
ここでは、$\sigma_k = \sigma$はすべて同じであるとする。

$p = 1$の場合の線形判別分析

正規分布の確率密度関数:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

ここで、μ_kと$\sigma_k{}^2$はそれぞれk番目のクラスの平均と分散である。
ここでは、$\sigma_k = \sigma$はすべて同じであるとする。
これをベイズの定理の式に代入すると、

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

というやや複雑な式が得られる。
ただし、 $p_k(x) = \Pr(Y = k|X = x)$とする。嬉しいことに、簡略化されたものもある。

判別関数

観測されたデータ$X = x$を$p_k(x)$が最大となるクラスに分類する。
$p_k(x)$の式の対数をとり、kに依存しない項を消すことにより、これは観測データを判別スコア

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

が最大となるクラスに分類することと同じであることがわかる。
なお、$\delta_k(x)$はxの線形関数である。

判別関数

観測されたデータ$X = x$を$p_k(x)$が最大となるクラスに分類する。
$p_k(x)$の式の対数をとり、kに依存しない項を消すことにより、これは観測データを判別スコア

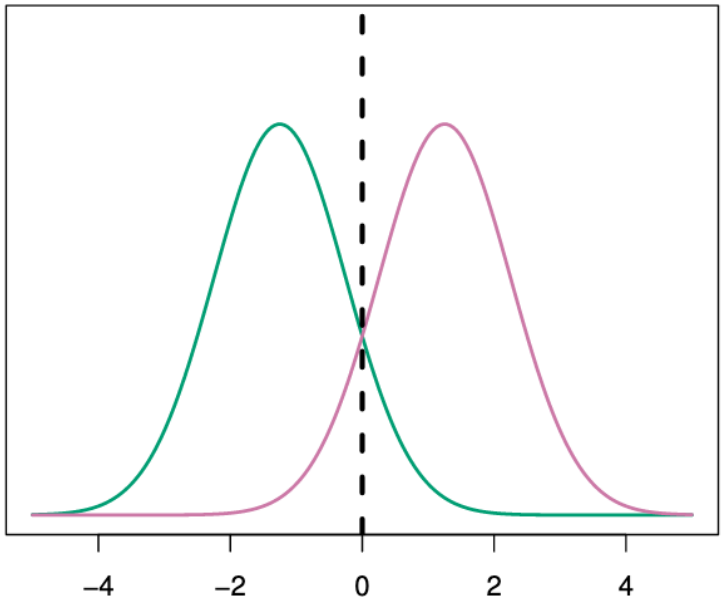
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

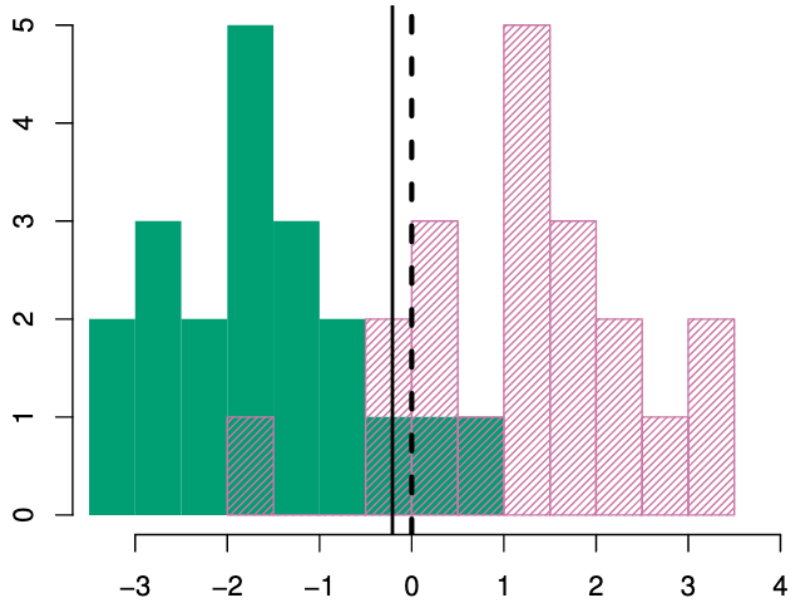
が最大となるクラスに分類することと同じであることがわかる。
なお、$\delta_k(x)$はxの線形関数である。

$K = 2$かつ$\pi_1 = \pi_2 = 0.5$の場合、ベイズの決定境界は

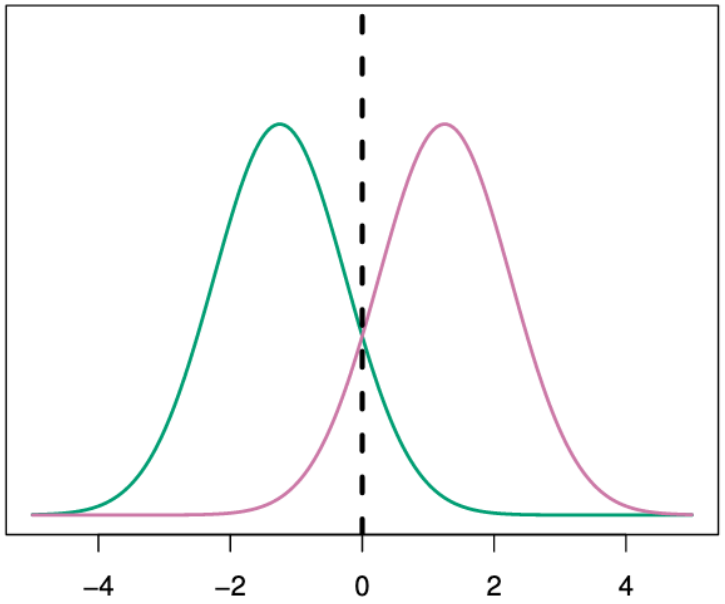
$$x = \frac{\mu_1 + \mu_2}{2}$$

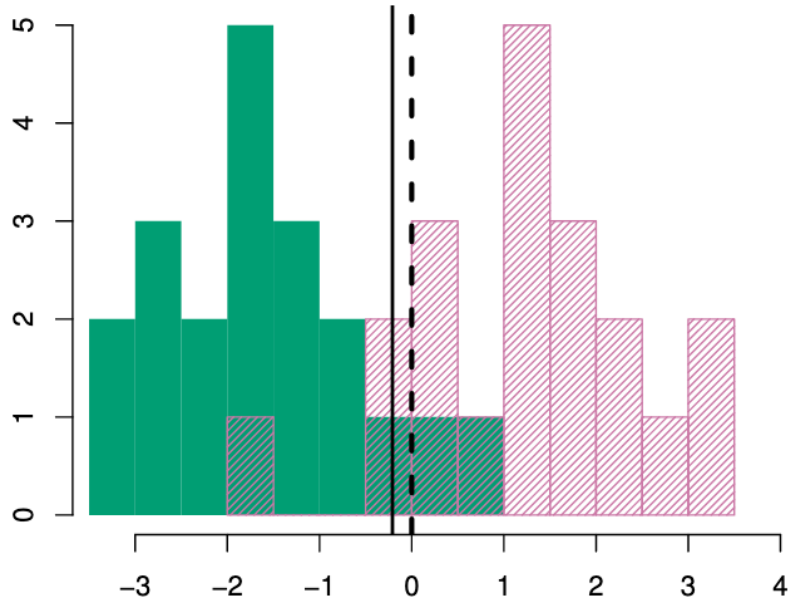
となる点に該当する。





例：$\mu_1 = -1.5, \mu_2 = 1.5, \pi_1 = \pi_2 = 0.5, \sigma^2 = 1$





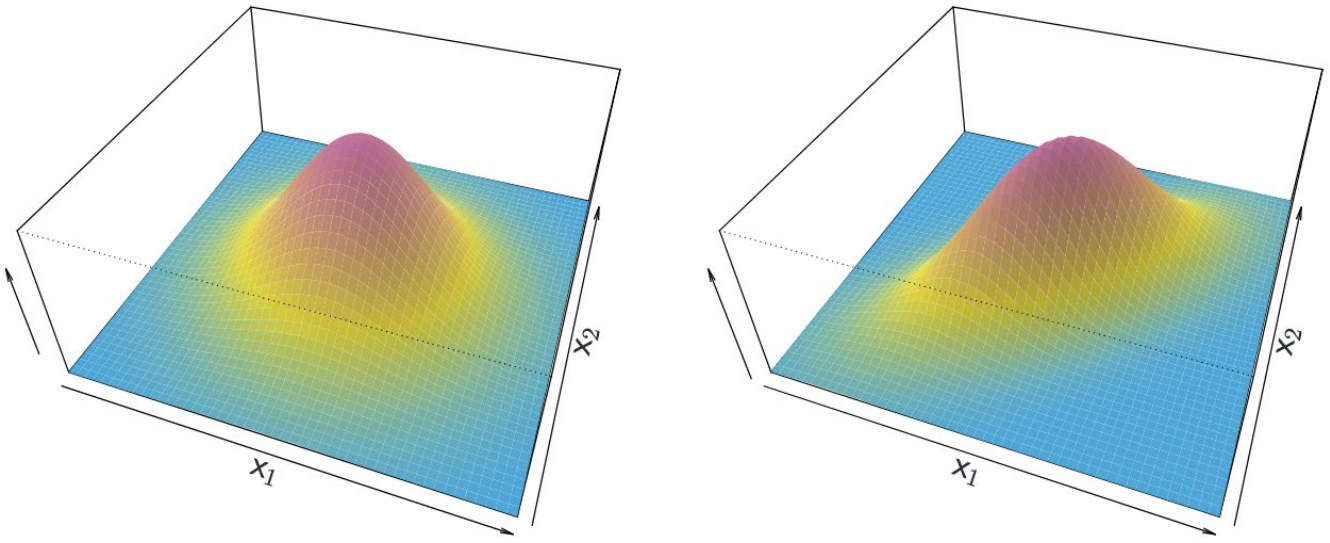
例：$\mu_1 = -1.5, \mu_2 = 1.5, \pi_1 = \pi_2 = 0.5, \sigma^2 = 1$
通常、これらのパラメータを知ることはなく、学習データがある。この場合、パラメータを推定し、判別スコアの式に代入すればよい。

パラメータの推定

$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:\, y_i = k} x_i \\
\hat{\sigma}^2 &= \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:\, y_i = k} (x_i - \hat{\mu}_k)^2 \\
&= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k^2
\end{aligned}
$$

ここで、$\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$は$k$番目のクラスに属する訓練データの標本分散である。

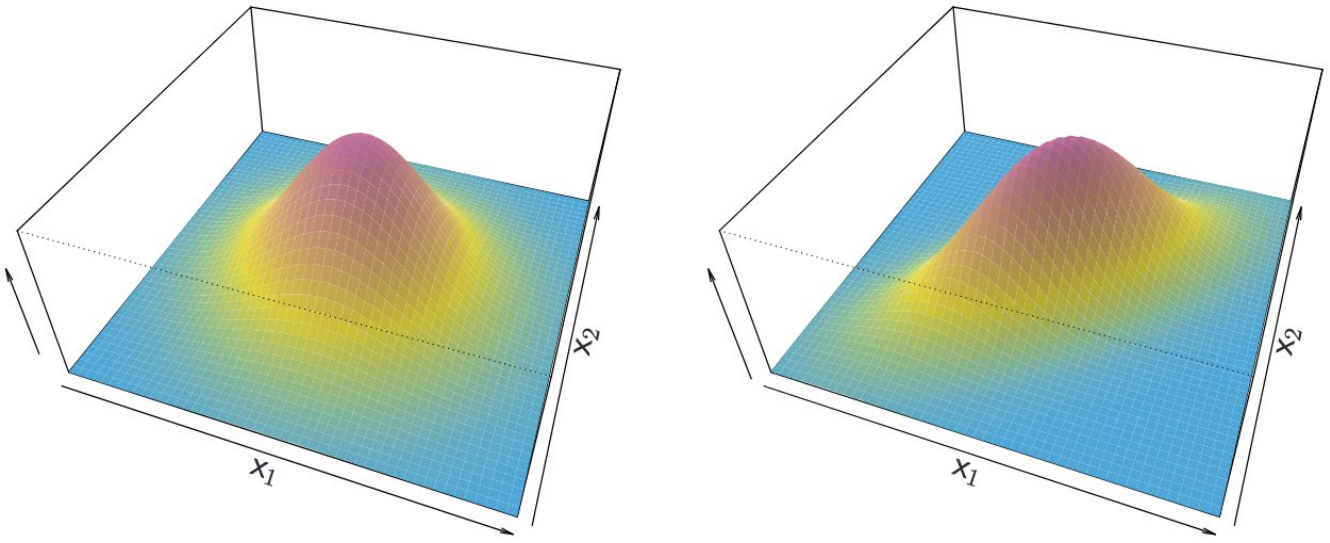
線形判別分析：$p > 1$の場合



多変量正規分布の確率密度関数：

$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)}$$

線形判別分析：$p > 1$の場合

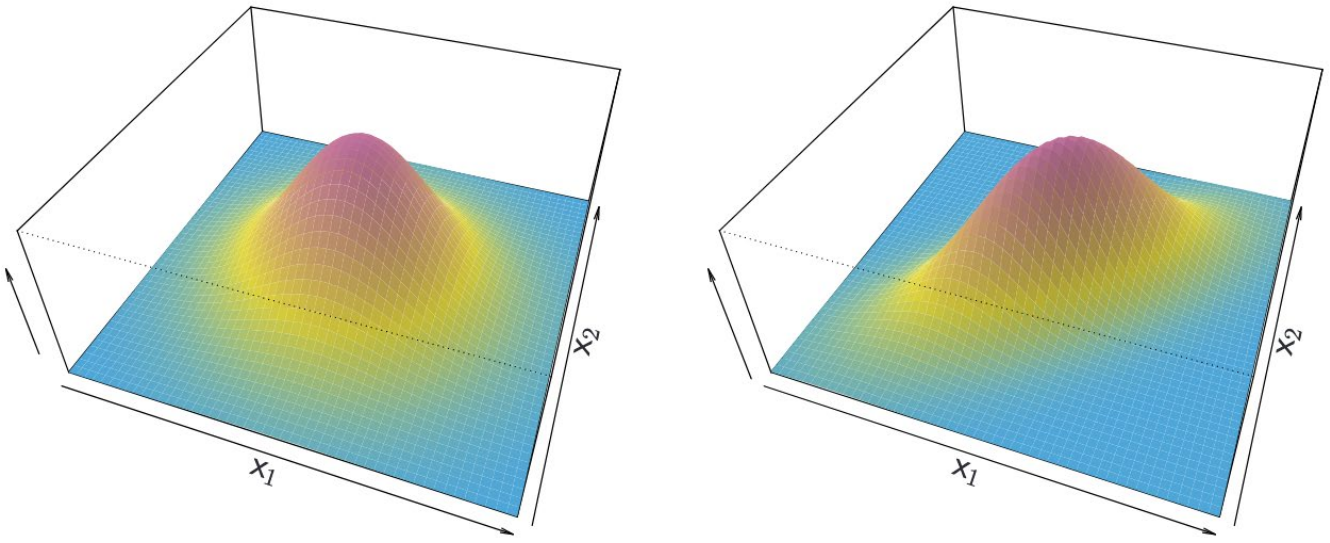


多変量正規分布の確率密度関数：

$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)}$$

判別関数：$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2} \mu_k^T \sum^{-1} \mu_k + \log \pi_k$

線形判別分析：$p > 1$の場合



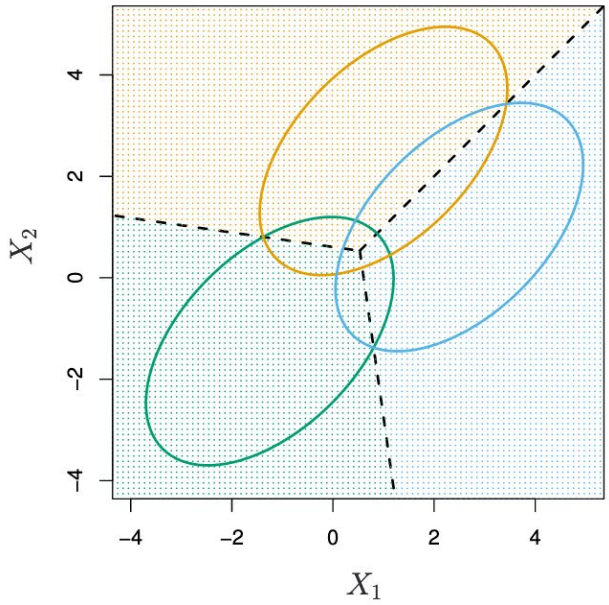
多変量正規分布の確率密度関数：

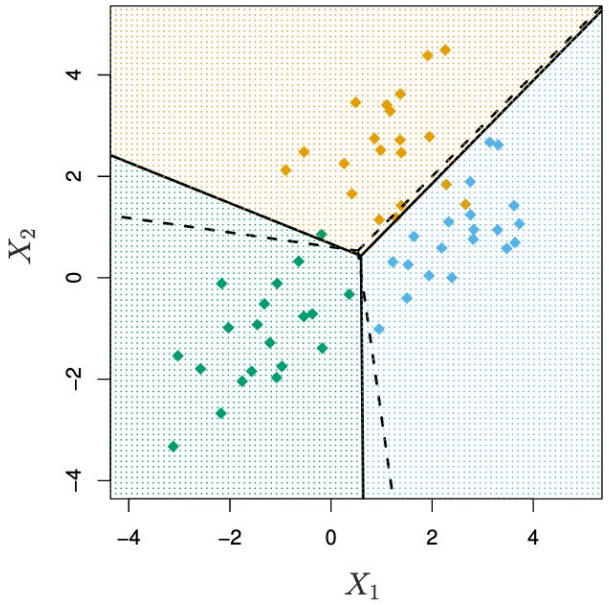
$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)}$$

判別関数：$\delta_k(x) = x^T \sum^{-1} \mu_k - \frac{1}{2}\mu_k^T \sum^{-1} \mu_k + \log \pi_k$

複雑な式にもかかわらず、$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \cdots + c_{kp}x_p$は線形関数である。

p=2, k=3の例





ここで、$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$である。

破線は、ベイズ決定境界と呼ばれるものである。これが分かっていれば、可能な分類法の中で最も誤分類率の少ない分類が実現できる。

あやめデータの例

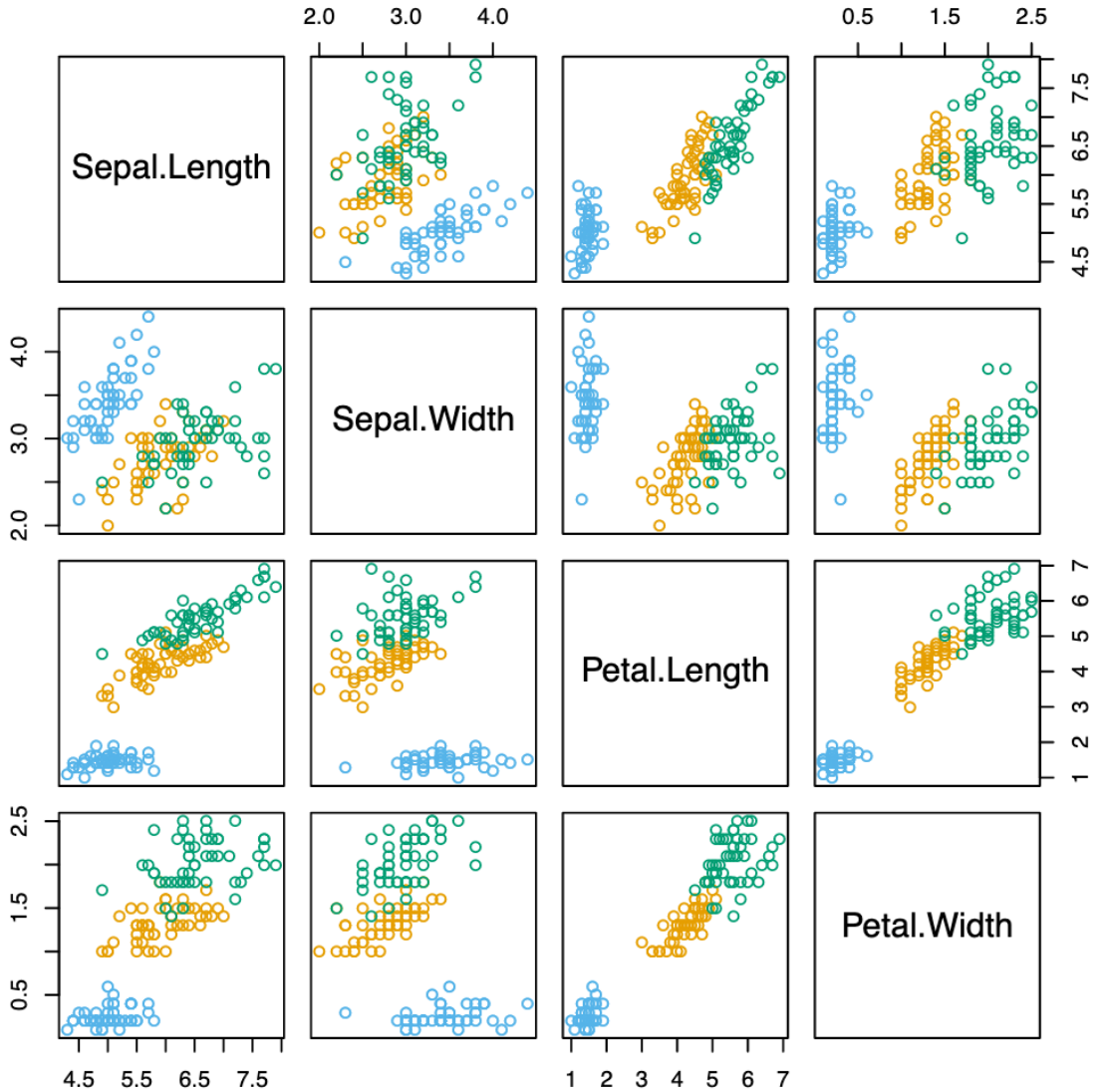
4つの特徴量（花びら・がく片の長さと幅）

あやめの種類:
3種類

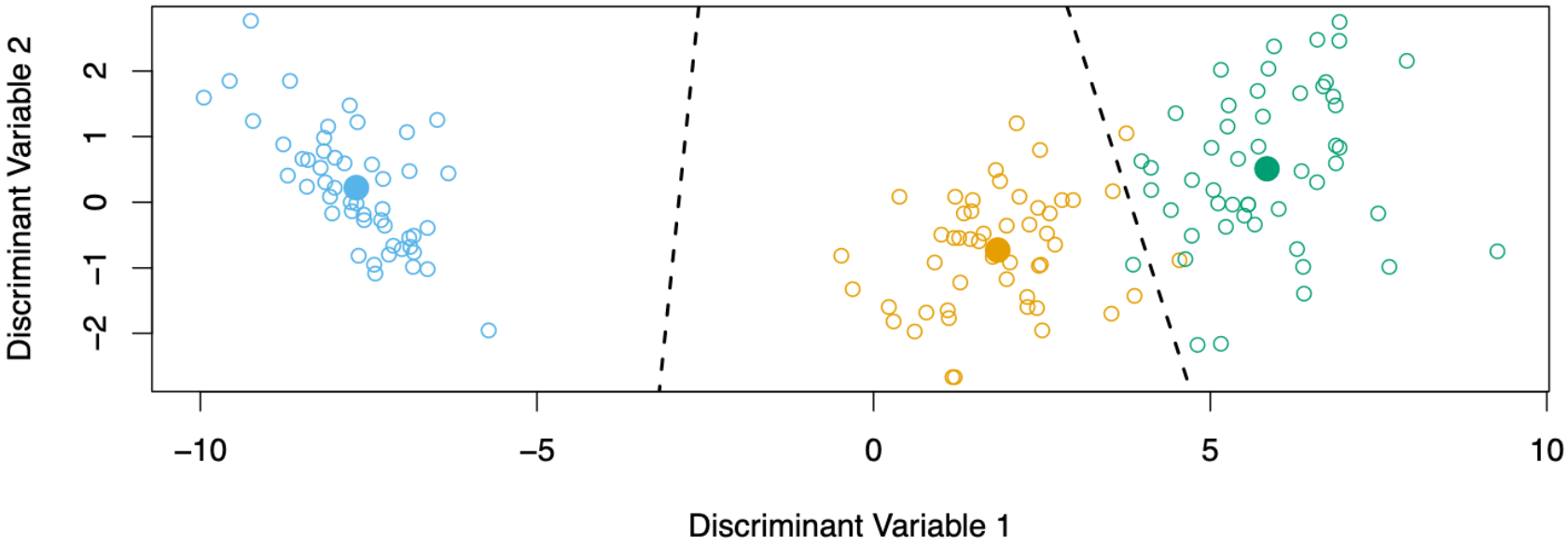
- Setosa
- Versicolor
- Virginica

それぞれ50件ずつある。

LDAにより、150個の訓練データのうち3つ以外のすべてが正しく分類された。

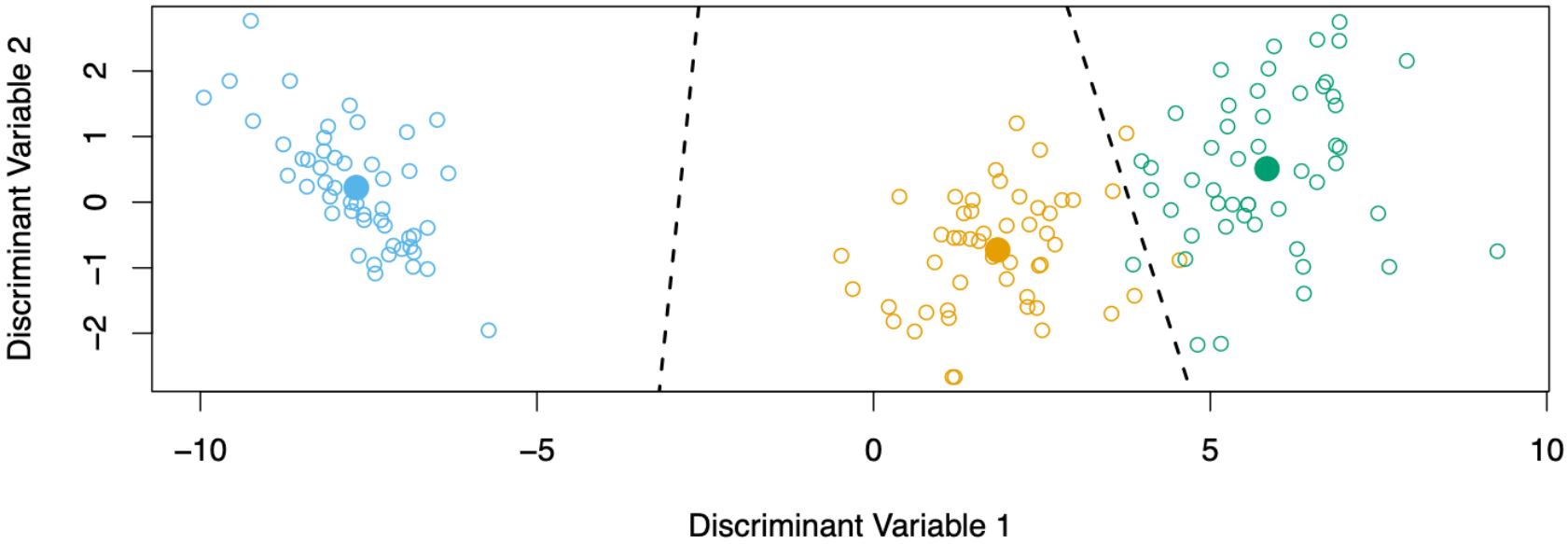


あやめデータの分析結果



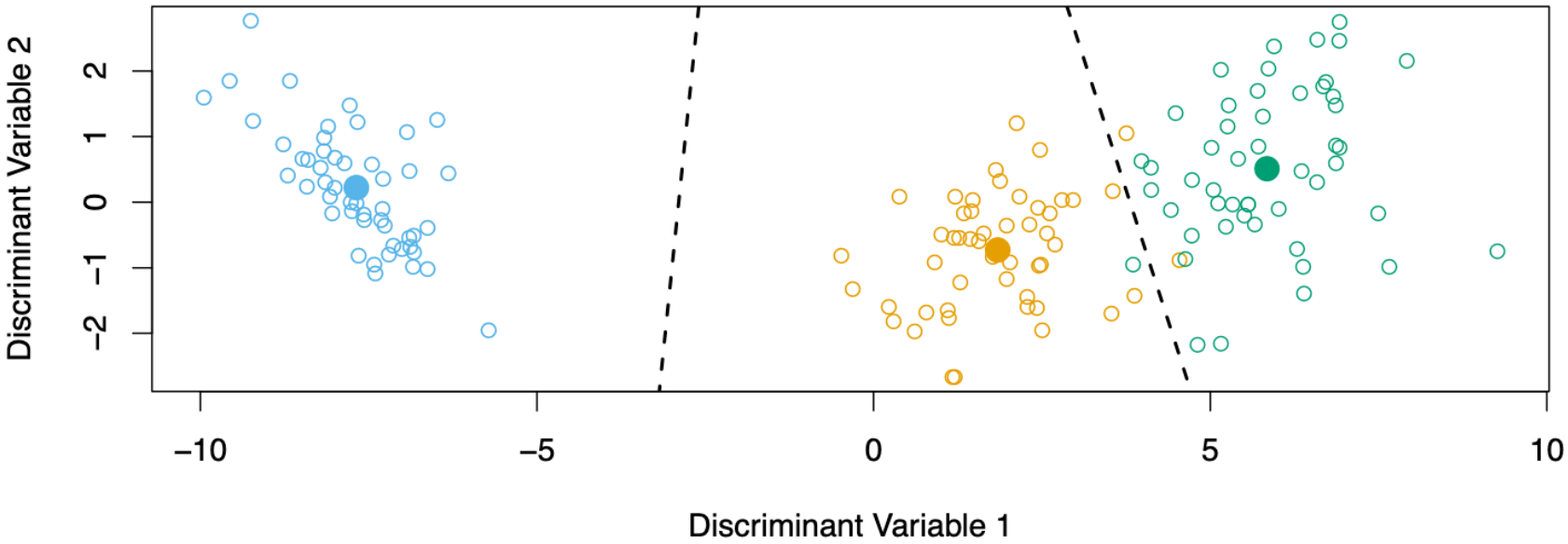
クラスがk個ある時、線形判別分析は$k-1$次元のプロットで可視化することができる。
なぜ？

あやめデータの分析結果



クラスがk個ある時、線形判別分析は$k-1$次元のプロットで可視化することができる。
なぜ？それは、線形判別分析は本質的に最も近い重心に応じて分類するからである。

あやめデータの分析結果



クラスがk個ある時、線形判別分析は$k-1$次元のプロットで可視化することができる。
なぜ？それは、線形判別分析は本質的に最も近い重心に応じて分類するからである。
$k > 3$の場合でも、判別ルールを可視化するために「最適」な2次元平面を視覚化することができる。

$\delta_k(x)$から確率へ

$\hat{\delta}_k(x)$の推定値が得られたら、これを用いてクラスkに属する確率を計算できる。

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

つまり、最も値の大きい$\hat{\delta}_k(x)$に応じて分類を行うということは、$\widehat{\Pr}(Y = k|X = x)$が最大となるクラスに分類することである。

$\delta_k(x)$から確率へ

$\hat{\delta}_k(x)$の推定値が得られたら、これを用いてクラスkに属する確率を計算できる。

$$\widehat{\Pr}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

つまり、最も値の大きい$\hat{\delta}_k(x)$に応じて分類を行うということは、$\widehat{\Pr}(Y = k|X = x)$が最大となるクラスに分類することである。
$k = 2$の時、$\widehat{\Pr}(Y = 2|X = x) \geq 0.5$ならばクラス2に分類する。
それ以外はクラス1に分類する。

LDAによるクレジットデータの分類結果

		実際の債務の状況		
		履行	不履行	合計
予測した債務の状況	履行	9644	252	9896
	不履行	23	81	104
	合計	9667	333	10000

10000個の訓練データに当てはめたLDAモデルの誤分類率は

$$\frac{23 + 252}{10000} \times 100\% = 2.75\%$$

となる。

いくつかの注意点：

・これは学習誤差であり、過学習している可能性がある。

LDAによるクレジットデータの分類結果

		実際の債務の状況		
		履行	不履行	合計
予測した債務の状況	履行	9644	252	9896
	不履行	23	81	104
	合計	9667	333	10000

10000個の訓練データに当てはめたLDAモデルの誤分類率は

$$\frac{23 + 252}{10000} \times 100\% = 2.75\%$$

となる。

いくつかの注意点：

・これは学習誤差であり、過学習している可能性がある。ただし、ここではn=10000, p=2なので、あまり心配する必要はない。

LDAによるクレジットデータの分類結果

		実際の債務の状況		
		履行	不履行	合計
予測した債務の状況	履行	9644	252	9896
	不履行	23	81	104
	合計	9667	333	10000

10000個の訓練データに当てはめたLDAモデルの誤分類率は

$$\frac{23+252}{10000}\times 100\% = 2.75\%$$

となる。

いくつかの注意点：

・これは学習誤差であり、過学習している可能性がある。ただし、ここではn=10000, p=2なので、あまり心配する必要はない。

・訓練データにおいて債務不履行に陥るのは3.33%なので、どの顧客も債務履行に属するような分類器では誤分類は3.33%しかない。

LDAによるクレジットデータの分類結果

		実際の債務の状況		
		履行	不履行	合計
予測した債務の状況	履行	9644	252	9896
	不履行	23	81	104
	合計	9667	333	10000

10000個の訓練データに当てはめたLDAモデルの誤分類率は

$$\frac{23+252}{10000}\times 100\% = 2.75\%$$

となる。

いくつかの注意点：

・これは学習誤差であり、過学習している可能性がある。ただし、ここではn=10000, p=2なので、あまり心配する必要はない。

・訓練データにおいて債務不履行に陥るのは3.33%なので、どの顧客も債務履行に属するような分類器では誤分類は3.33%しかない。

・LDAでは、9667人の債務履行の人のうち23人が誤って分類された（誤分類率：23/9667=0.2%）。333人の債務不履行に陥った人のうち、252人が誤って分類された（誤分類率：252/333=75.7%）。

誤分類率の種類

偽陽性率：本当は陰性であるものの、分類が誤って陽性と判断される割合。例では0.2%である。
偽陰性率：本当は陽性であるものの、分類が誤って陰性と判断される割合。例では75.5%である。

先程の表では、もし

$$\Pr(Default = \text{債務不履行} | \text{債務残高}, \text{学生}) \geq 0.5$$

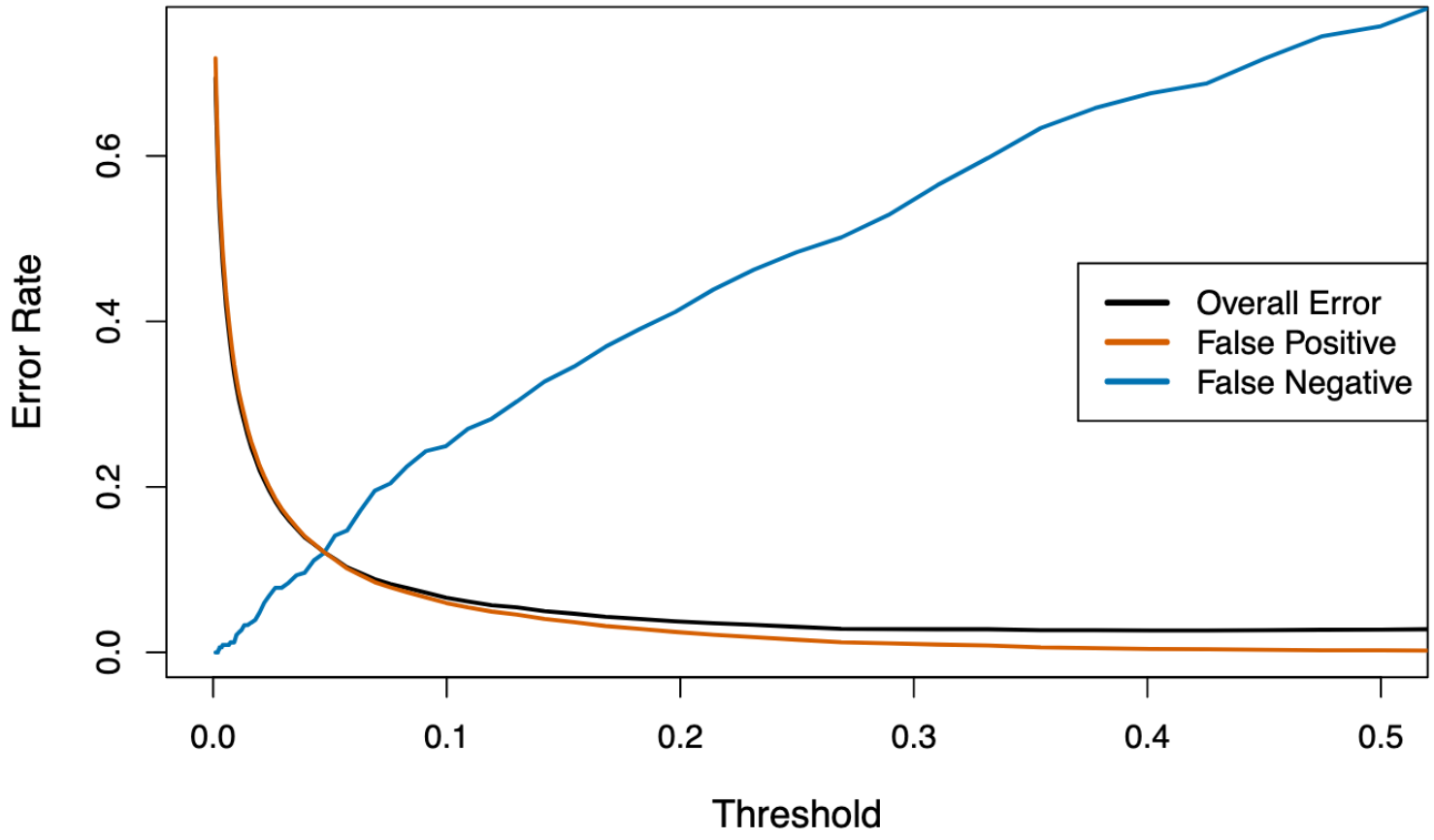
であれば、債務不履行のクラスに分類することになる。

偽陽性率と偽陰性率を制御するには、

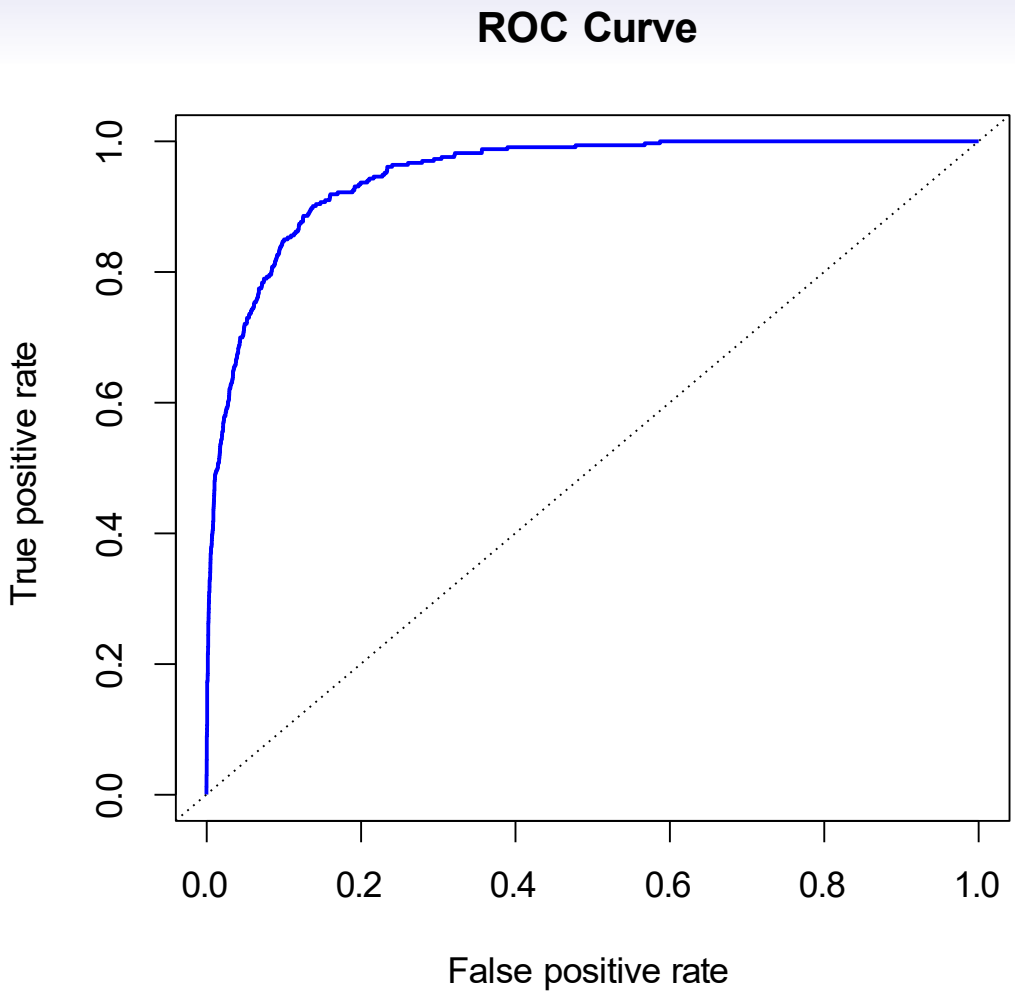
$$\Pr(Default = \text{債務不履行} | \text{債務残高}, \text{学生}) \geq \text{閾値}$$

の閾値を[0,1]の間の値で変更すればよい。

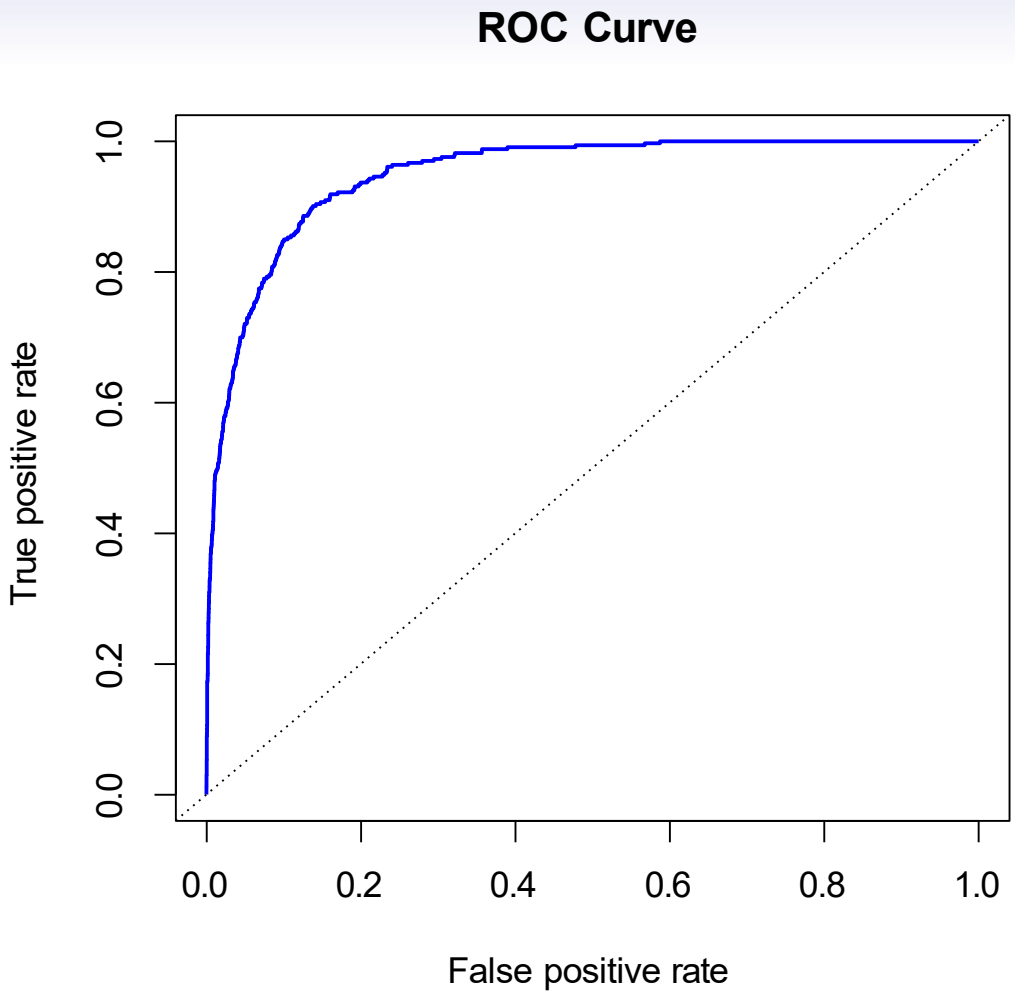
閾値を変化させる



偽陰性率を下げるために、閾値を0.1以下にするとよかろう。



ROC曲線は、偽陽性率と偽陰性率を同時に表現するものである。



ROC曲線は、偽陽性率と偽陰性率を同時に表現するものである。全体的な性能を表すものとして、AUC（ROC曲線の下方面積）がある。AUCは大きいほど良いことになる。

判別分析のその他の表現

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$f_k(x)$が多変量正規分布の確率密度関数で、かつ全てのクラスに共通の分散共分散行列Σをもつとき、線形判別分析になる。

$f_k(x)$の形式を変えることで、異なる分類器が得られる。

- 各クラスでの観測データは、クラスごとに異なる分散共分散行列Σ_kをもつ多変量正規分布に従う場合、2次判別分析になる。

判別分析のその他の表現

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$f_k(x)$が多変量正規分布の確率密度関数で、かつ全てのクラスに共通の分散共分散行列Σをもつとき、線形判別分析になる。
$f_k(x)$の形式を変えることで、異なる分類器が得られる。

- 各クラスでの観測データは、クラスごとに異なる分散共分散行列Σ_kをもつ多変量正規分布に従う場合、2次判別分析になる。
- $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$という条件付き独立を仮定したモデルを各クラスに適用すると、ナイーブベイズになる。

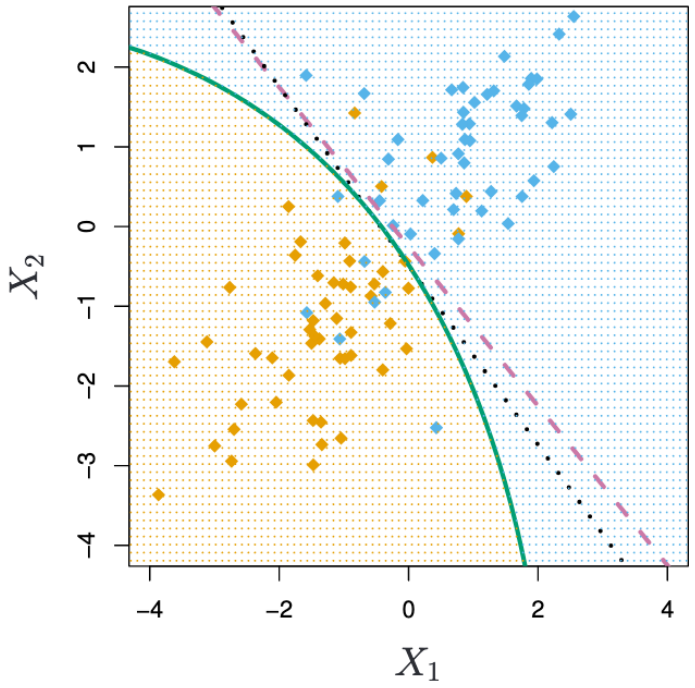
判別分析のその他の表現

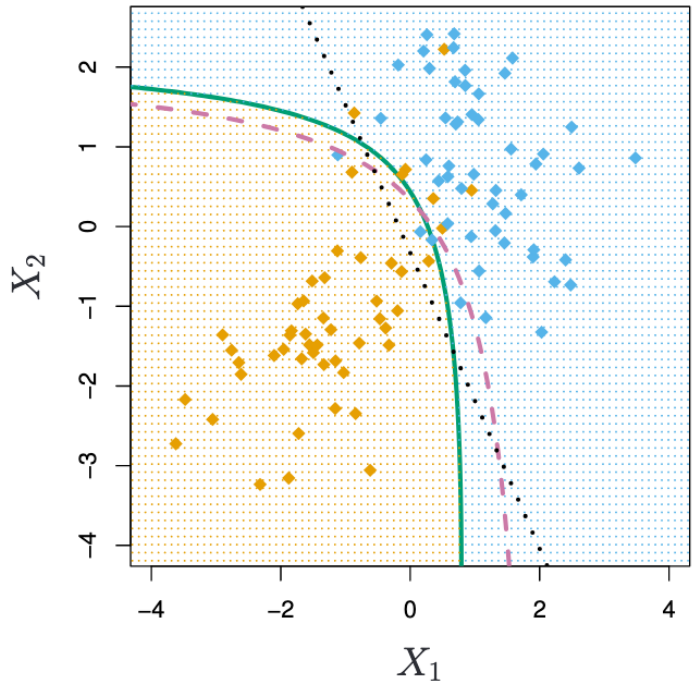
$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

$f_k(x)$が多変量正規分布の確率密度関数で、かつ全てのクラスに共通の分散共分散行列Σをもつとき、線形判別分析になる。
$f_k(x)$の形式を変えることで、異なる分類器が得られる。

- 各クラスでの観測データは、クラスごとに異なる分散共分散行列Σ_kをもつ多変量正規分布に従う場合、2次判別分析になる。
- $f_k(x) = \prod_{j=1}^{p} f_{jk}(x_j)$という条件付き独立を仮定したモデルを各クラスに適用すると、ナイーブベイズになる。
- そのほか、$f_k(x)$に対してノンパラメトリックな手法を適用する方法もある。

2次判別分析





$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) + \log\pi_k - \frac{1}{2}\log|\Sigma_k|$$

$\sum_k$が異なるため、この式において2次の項が重要である。

ナイーブベイズ

各クラスでの特徴量が互いに独立であることを仮定する。
この手法は、pが大きい場合に有効である。

- ガウス型ナイーブベイズは、各Σ_kが対角であると仮定する。

$$\begin{aligned}\delta_k(x) &\propto \log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] \\ &= -\frac{1}{2}\sum_{j=1}^{p}\left[\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + \log\sigma_{kj}^2\right] + \log\pi_k\end{aligned}$$

- この手法は、混合特徴ベクトル（量的と質的）を扱うことができる。X_jが質的変数の場合は、$f_{kj}(x_j)$を確率質量関数と変更すればよい。

強い仮定にも関わらず、ナイーブベイズはしばしば良質な分類結果を出す。

ロジスティック回帰と線形判別分析の比較

2値分類の問題では、

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p$$

が成立する。つまり、LDAはロジスティック回帰と同じ形をしている。
両者の違いは、パラメータの推定方法にある。

ロジスティック回帰と線形判別分析の比較

2値分類の問題では、

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p$$

が成立する。つまり、LDAはロジスティック回帰と同じ形をしている。

両者の違いは、パラメータの推定方法にある。

- ロジスティック回帰は、条件付き尤度を使用している。これは、$\Pr(Y|X)$に基づくものであり、判別学習として知られている。

ロジスティック回帰と線形判別分析の比較

2値分類の問題では、

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p$$

が成立する。つまり、LDAはロジスティック回帰と同じ形をしている。

両者の違いは、パラメータの推定方法にある。

- ロジスティック回帰は、条件付き尤度を使用している。これは、$\Pr(Y|X)$に基づくものであり、判別学習として知られている。
- LDAは、$\Pr(X,Y)$に基づく完全尤度を使用している。これは、生成学習として知られている。

ロジスティック回帰と線形判別分析の比較

2値分類の問題では、

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p$$

が成立する。つまり、LDAはロジスティック回帰と同じ形をしている。

両者の違いは、パラメータの推定方法にある。

- ロジスティック回帰は、条件付き尤度を使用している。これは、$\Pr(Y|X)$に基づくものであり、判別学習として知られている。
- LDAは、$\Pr(X,Y)$に基づく完全尤度を使用している。これは、生成学習として知られている。
- これらの違いにもかかわらず、実際には、結果はしばしば非常に類似している。

ロジスティック回帰と線形判別分析の比較

2値分類の問題では、

$$\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x_1 + \cdots + c_p x_p$$

が成立する。つまり、LDAはロジスティック回帰と同じ形をしている。

両者の違いは、パラメータの推定方法にある。

- ロジスティック回帰は、条件付き尤度を使用している。これは、$\Pr(Y|X)$に基づくものであり、判別学習として知られている。
- LDAは、$\Pr(X,Y)$に基づく完全尤度を使用している。これは、生成学習として知られている。
- これらの違いにもかかわらず、実際には、結果はしばしば非常に類似している。

注：ロジスティック回帰は、モデルに二次項を明示的に含めることで、QDAのような二次境界もフィットできる。

まとめ

- $k = 2$のときによく、ロジスティック回帰は分類のために使われる。
- 線形判別分析は、以下の場合において有効と考えられる。
  nが小さい
  クラスがよく分類されている
  正規分布の仮定が妥当である
  $k > 2$のとき
- ナイーブベイズは、pが非常に大きい場合に有効である。
- ロジスティック回帰、LDA、KNNの違いについて、Section4.5を参照してください。

第5章 リサンプリング法
-Resampling methods-

- 訓練誤差と予測誤差
- リサンプリング
- ホールドアウト法
- K分割交差検証法
- ブートストラップ法
- プレバリデーション

第5章 リサンプリング法
-Resampling methods-

- この章では、リサンプリングの手法である、交差検証とブートストラップについて論じる。

交差検証とブートストラップ

- この章では、リサンプリングの手法である、交差検証とブートストラップについて論じる。
- これらの方法は、訓練データから標本を抽出し、関心のあるモデルに適用し、それを繰り返すことでモデルに関する新たな情報を得ることができる。

交差検証とブートストラップ

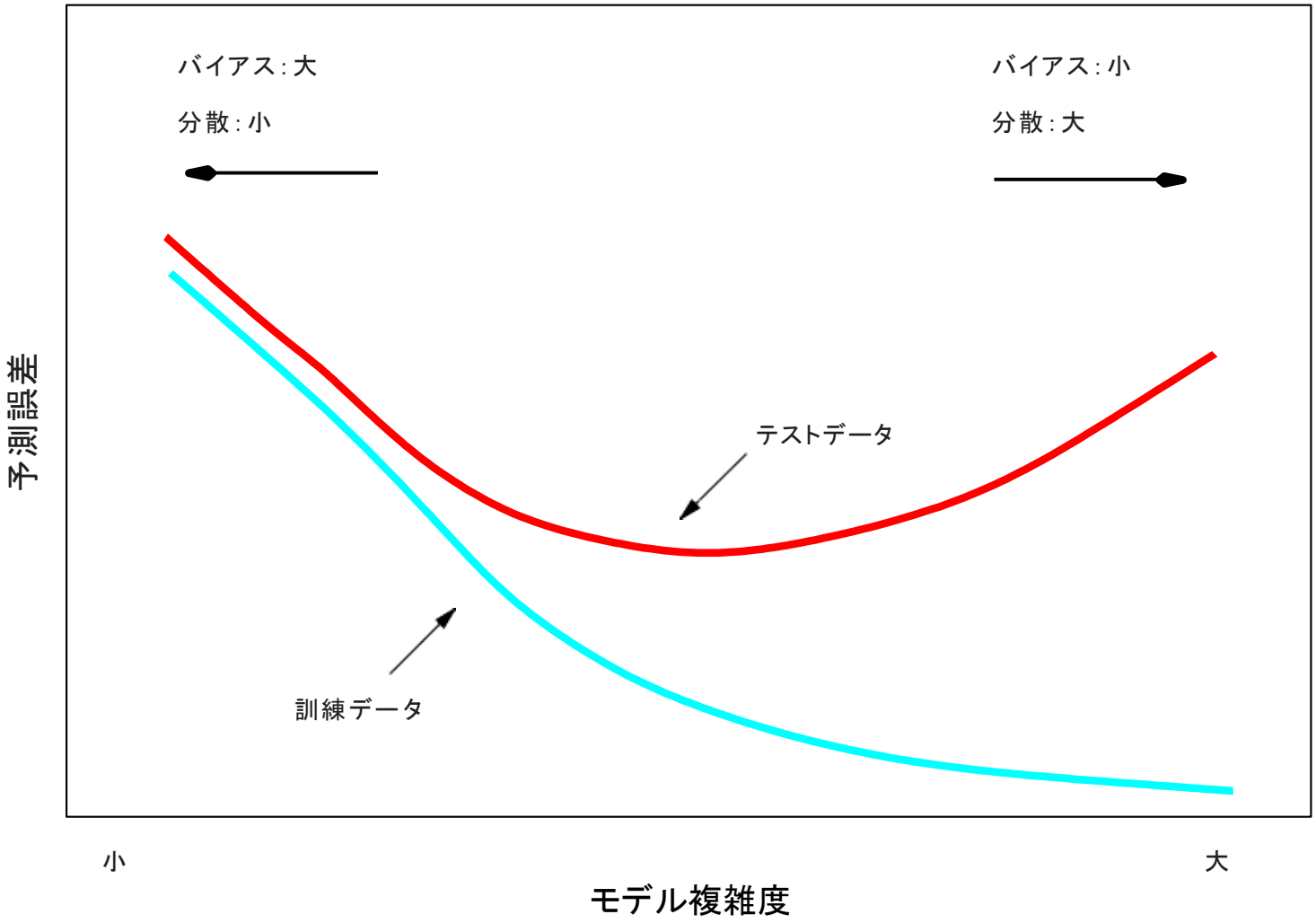
- この章では、リサンプリングの手法である、交差検証とブートストラップについて論じる。
- これらの方法は、訓練データから標本を抽出し、関心のあるモデルに適用し、それを繰り返すことでモデルに関する新たな情報を得ることができる。
- 例えば、テストデータによる予測誤差の推定値や、推定したパラメータの標準偏差やバイアスを求めることができる。

訓練誤差　versus テスト誤差

以下は、テスト誤差と訓練誤差の違いについて再掲する。

- テスト誤差とは、新たな観測データに対して、統計的学習方法を用いて予測した場合の平均的な誤差を指す。この新たな観測データとは、学習時には使用されなかったデータのことである。
- 一方、訓練誤差は、学習に使用されたデータに対して統計的学習方法を適用することによって簡単に計算できる。
- しかし、通常、訓練誤差はテスト誤差と大きく異なり、特に前者は後者よりもかなり小さい値になることがある。

訓練データ　versus テストデータ



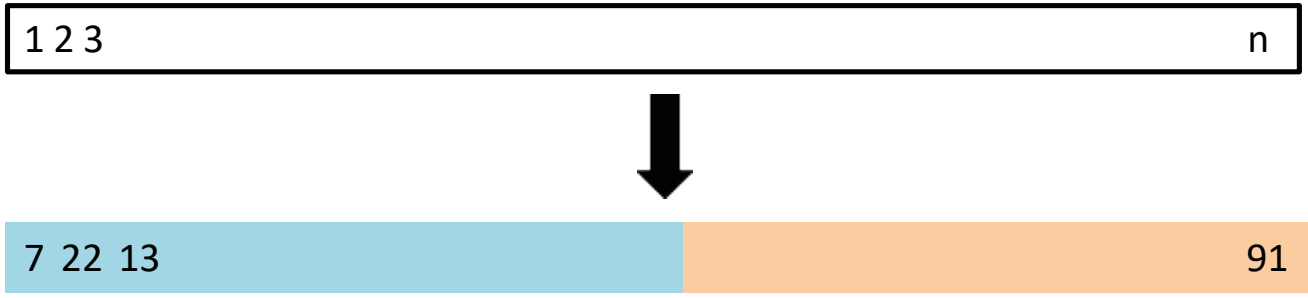
予測誤差の推定について

- 最良の解決策：大規模なテストセット。しばしば利用不可能
- いくつかの方法は、訓練誤差率を数学的に調整してテスト誤差率を推定する。これには、Cp統計量、AIC、BICが含まれる。
- ここでは、モデルの当てはめを行う際に訓練データのうちいくつかを取り置きして、その取り置きしたデータに統計的学習法を適用することにより、テスト誤差を推定する方法を考える。

ホールドアウト検証

- このアプローチでは、観測データをランダムに2つのデータセットに分割する：訓練データと検証データ。
- モデルは訓練データに適合させ、適合したモデルは、検証データの観測値に対する応答を予測するために使用される。
- 結果として得られる検証データの誤差は、テスト誤差の推定値となる。これは、数量的な応答の場合にはMSEを使用し、質的な（離散的な）応答の場合には分類誤差率を使用して評価される。

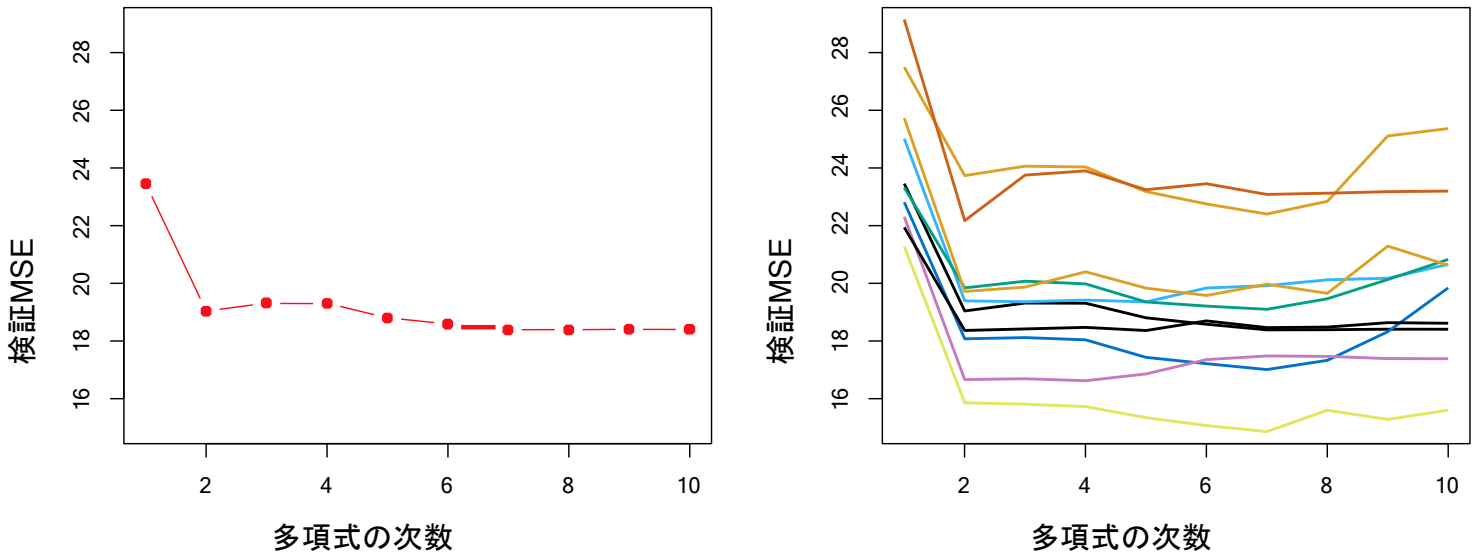
ホールドアウト検証



観測データをランダムに訓練データと検証データに分割する。

例: automobile data

- 線形回帰と多項式回帰を比較する。
- 392個の観測データを196個の訓練データと196個の検証データに分割する。



左：1回の分割　右：複数回の分割

ホールドアウト検証の問題点

- ホールドアウト検証によるテスト誤差の推定値は、どのデータが訓練データに含まれるか、または検証データに含まれるかによってかなりばらつきがあることに注意が必要である。
- ホールドアウト検証では、訓練データに含まれる観測値のみがモデルの当てはめに使用され、検証データに含まれる観測値は使用されない。
- これにより、検証誤差は、すべての観測データを用いた場合のモデルのテスト誤差を過小評価する可能性があることを示唆している。

ホールドアウト検証の問題点

- ホールドアウト検証によるテスト誤差の推定値は、どのデータが訓練データに含まれるか、または検証データに含まれるかによってかなりばらつきがあることに注意が必要である。
- ホールドアウト検証では、訓練データに含まれる観測値のみがモデルの当てはめに使用され、検証データに含まれる観測値は使用されない。
- これにより、検証誤差は、すべての観測データを用いた場合のモデルのテスト誤差を過小評価する可能性があることを示唆している。なぜ？

K分割交差検証

- テスト誤差を推定するために広く使用される手法。
- 推定結果は、最適なモデルを選択したり、最終的に選択されたモデルのテスト誤差を知るために使用することができる。
- 観測データをK個のほぼ同じサイズのグループにランダムに分割し、最初のグループを検証に使用し、残りのK-1グループでモデルを当てはめる。その後、残されたグループで平均二乗誤差（MSE）を計算する 。
- これをK回繰り返し、毎回異なるグループを検証データとして使用する。最終的な結果はK回の推定結果を結合したものとなる。

K分割交差検証

データをおよそ同じサイズのk個のグループに分割する（k＝5）

1	2	3	4	5
検証	訓練	訓練	訓練	訓練

K分割交差検証

- K個のグループをそれぞれ$C_1, C_2, \ldots, C_K$とする。ここでC_kはk番目のグループの観測データを表す。グループkのデータ数をn_kとすると、$n_k = n/K$である。

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{K} MSE_k$$

を計算する。ただし、$MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$とする。$\hat{y}_i$が$i$番目の観測値の推定値である。

K分割交差検証

- K個のグループをそれぞれ$C_1, C_2, \ldots, C_K$とする。ここでC_kはk番目のグループの観測データを表す。グループkのデータ数をn_kとすると、$n_k = n/K$である。

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{K} MSE_k$$

を計算する。ただし、$MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$とする。$\hat{y}_i$が$i$番目の観測値の推定値である。

- $K = n$の場合は、K分割交差検証がn分割交差検証となり、1つ抜き交差検証（LOOCV）の推定量と一致する。

良い特殊例！

- 最小2乗法による線形または多項式回帰では、以下の式

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

により、LOOCVをただ1つのモデルを当てはめるのと同じ計算量で行うことができる。ここに、$\hat{y}_i$はi番目の観測データに対する最小2乗推定量の推定値である。また、h_iはてこ比である。これは通常のMSEのようである。違いはi番目の残差を$1 - h_i$で割っていることである。

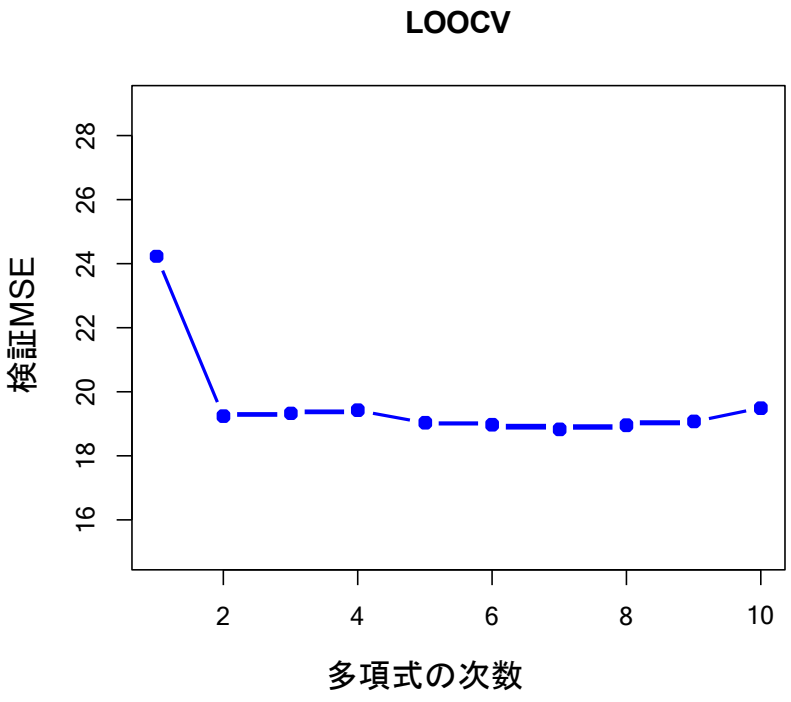
良い特殊例！

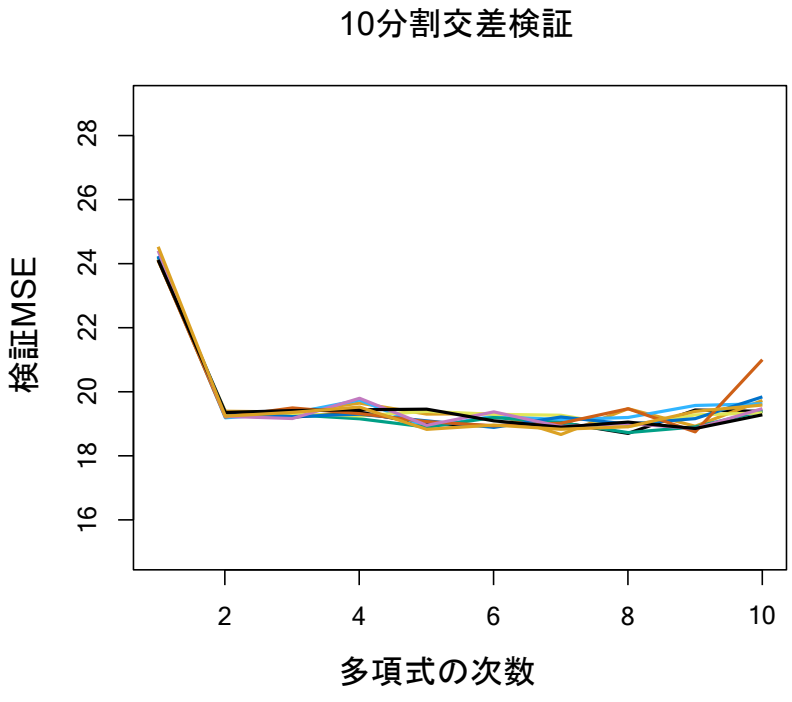
- 最小2乗法による線形または多項式回帰では、以下の式

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

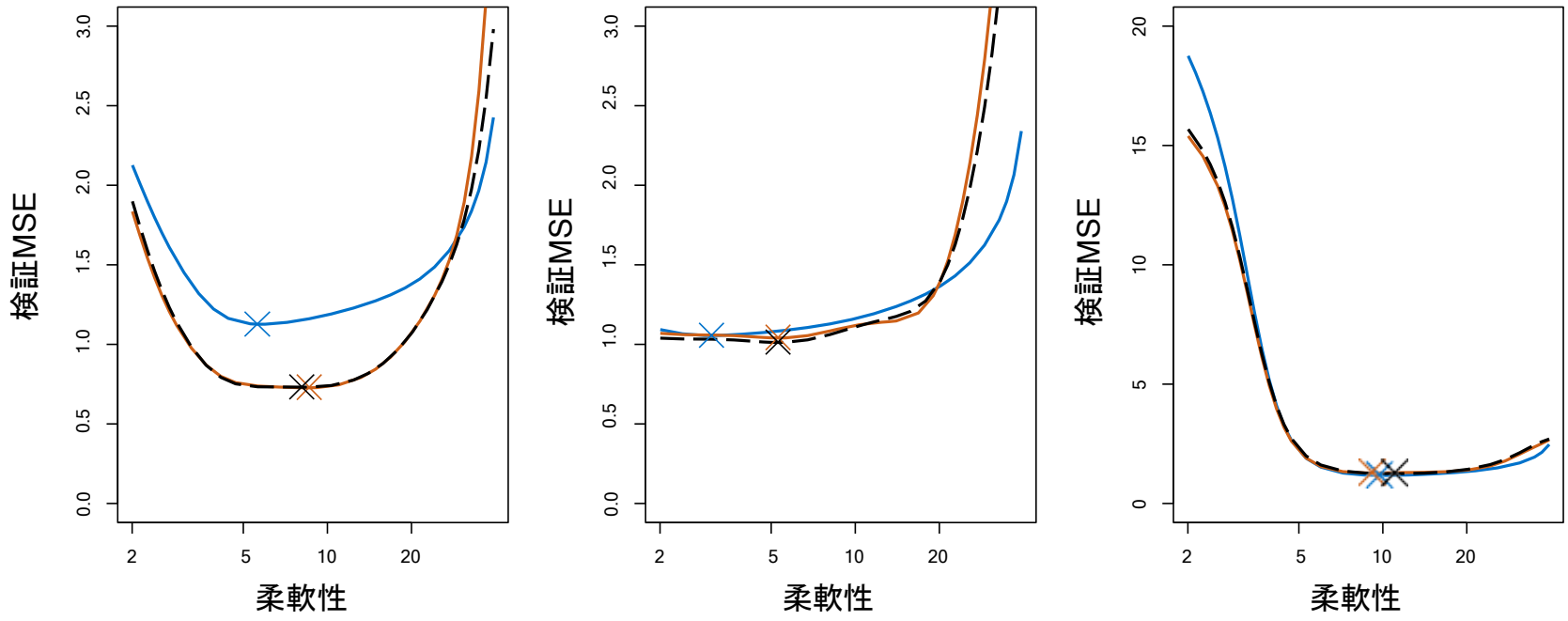
により、LOOCVをただ1つのモデルを当てはめるのと同じ計算量で行うことができる。ここに、$\hat{y}_i$はi番目の観測データに対する最小2乗推定量の推定値である。また、h_iはてこ比である。これは通常のMSEのようである。違いはi番目の残差を$1 - h_i$で割っていることである。
- K分割交差検証においては、$k = 5$や$k = 10$がよく使われる。

Auto data


LOOCV
検証MSE
16
18
20
22
24
26
28
2
4
6
8
10
多項式の次数


10分割交差検証
検証MSE
16
18
20
22
24
26
28
2
4
6
8
10
多項式の次数

シミュレーションデータの真のテストMSEとテストMSEの推定値



交差検証におけるその他の問題点

- それぞれの訓練データセットが、オリジナルの訓練データセットの (K - 1)/K しかないため、予測誤差の推定値は通常上方バイアスがかかる。

交差検証におけるその他の問題点

- それぞれの訓練データセットが、オリジナルの訓練データセットの (K - 1)/K しかないため、予測誤差の推定値は通常上方バイアスがかかる。なぜ？

交差検証におけるその他の問題点

- それぞれの訓練データセットが、オリジナルの訓練データセットの (K - 1)/K しかないため、予測誤差の推定値は通常上方バイアスがかかる。なぜ？
- このバイアスはK = n（LOOCV）の場合に最小化されるが、前述のようにこの推定値は高い分散を持つ。
- K = 5または10は、このバイアスと分散のトレードオフの良い妥協点を提供する。

分類における交差検証

- K個のグループをそれぞれ$C_1, C_2, \ldots, C_K$とする。ここでC_kはk番目のグループの観測データを表す。グループkのデータ数をn_kとすると、$n_k = n/K$である。

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{K} Err_k$$

を計算する。ただし、$Err_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$とする。

- CV_kの標準偏差の推定値は以下の式で計算できる。

$$\widehat{SE}(CV_k) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \frac{(Err_k - \overline{Err_k})^2}{K-1}}$$

- これは有用な推定量だが、厳密に言えば有効ではない。

分類における交差検証

- K個のグループをそれぞれ$C_1, C_2, \ldots, C_K$とする。ここでC_kはk番目のグループの観測データを表す。グループkのデータ数をn_kとすると、$n_k = n/K$である。

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{K} Err_k$$

を計算する。ただし、$Err_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$とする。

- CV_kの標準偏差の推定値は以下の式で計算できる。

$$\widehat{SE}(CV_k) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \frac{(Err_k - \overline{Err_k})^2}{K-1}}$$

- これは有用な推定量だが、厳密に言えば有効ではない。
  なぜ？

交差検証：正しい方法と誤った方法

- 2値分類問題のための分類器を考えてみよう。
    1. 5000の予測変数と50のサンプルを用いて、クラスのラベルと最も相関の高い100の予測変数を見つけたとする。
    2. 次に、これらの100の予測変数だけを使用して、ロジスティック回帰などの分類器を適用する。

この分類器のテストデータをどのように評価するか？

交差検証：正しい方法と誤った方法

- 2値分類問題のための分類器を考えてみよう。
    1. 5000の予測変数と50のサンプルを用いて、クラスのラベルと最も相関の高い100の予測変数を見つけたとする。
    2. 次に、これらの100の予測変数だけを使用して、ロジスティック回帰などの分類器を適用する。

この分類器のテストデータをどのように評価するか？

ステップ1を忘れて、ステップ2で交差検証を適用できるのか？

NO！

- 訓練データのラベルは、ステップ1で既に使用されているため、交差検証のプロセスにも含める必要がある。
- クラスのラベルが結果とは独立しているリアルなデータをシミュレートすることは簡単であり、真のテスト誤差が50%であるにもかかわらず、ステップ1を無視したCVの推定値はゼロになる！

NO！

- 訓練データのラベルは、ステップ1で既に使用されているため、交差検証のプロセスにも含める必要がある。
- クラスのラベルが結果とは独立しているリアルなデータをシミュレートすることは簡単であり、真のテスト誤差が50%であるにもかかわらず、ステップ1を無視したCVの推定値はゼロになる！自分でやってみてください。

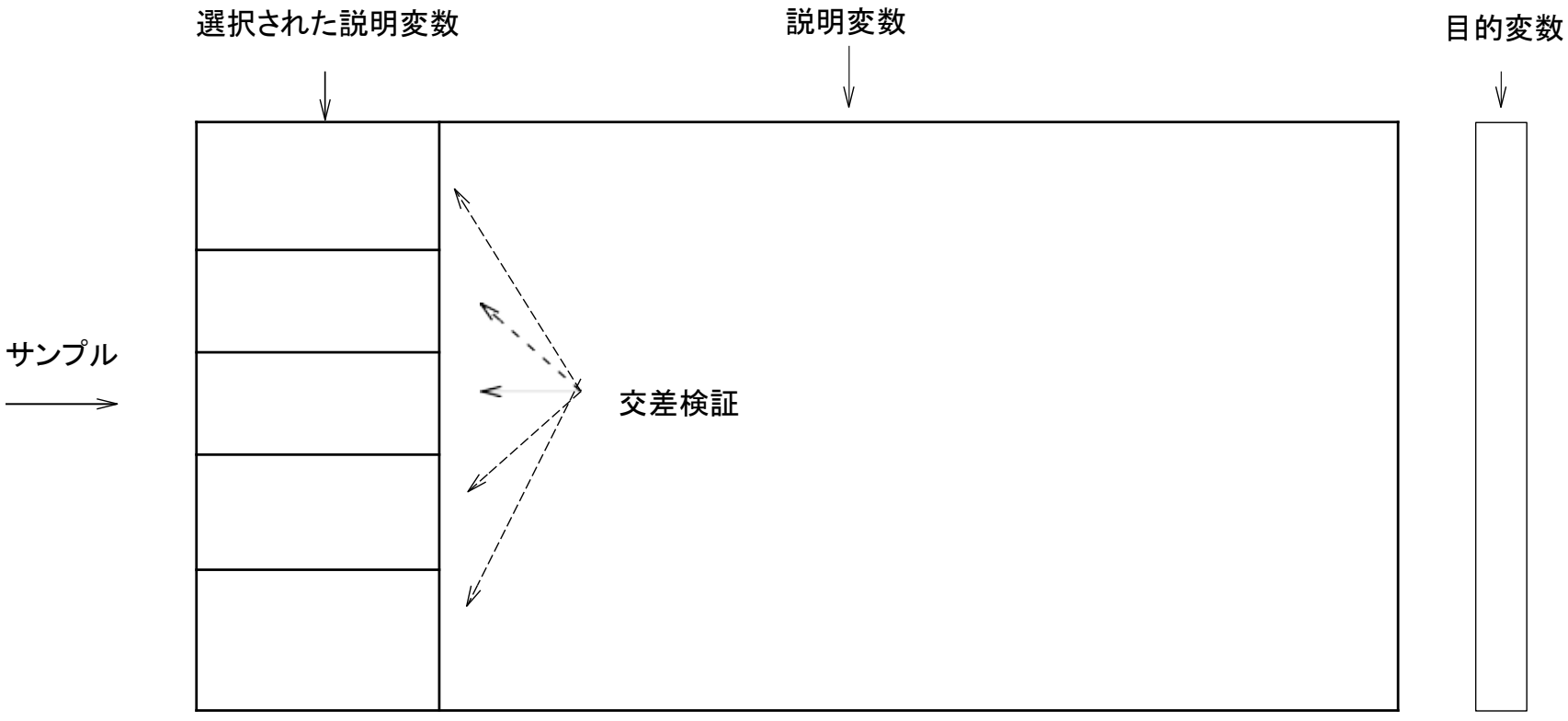
NO！

- 訓練データのラベルは、ステップ1で既に使用されているため、交差検証のプロセスにも含める必要がある。
- クラスのラベルが結果とは独立しているリアルなデータをシミュレートすることは簡単であり、真のテスト誤差が50%であるにもかかわらず、ステップ1を無視したCVの推定値はゼロになる！自分でやってみてください。
- 多くの有名なジャーナルの投稿論文でもこの誤ったやり方が見られる。

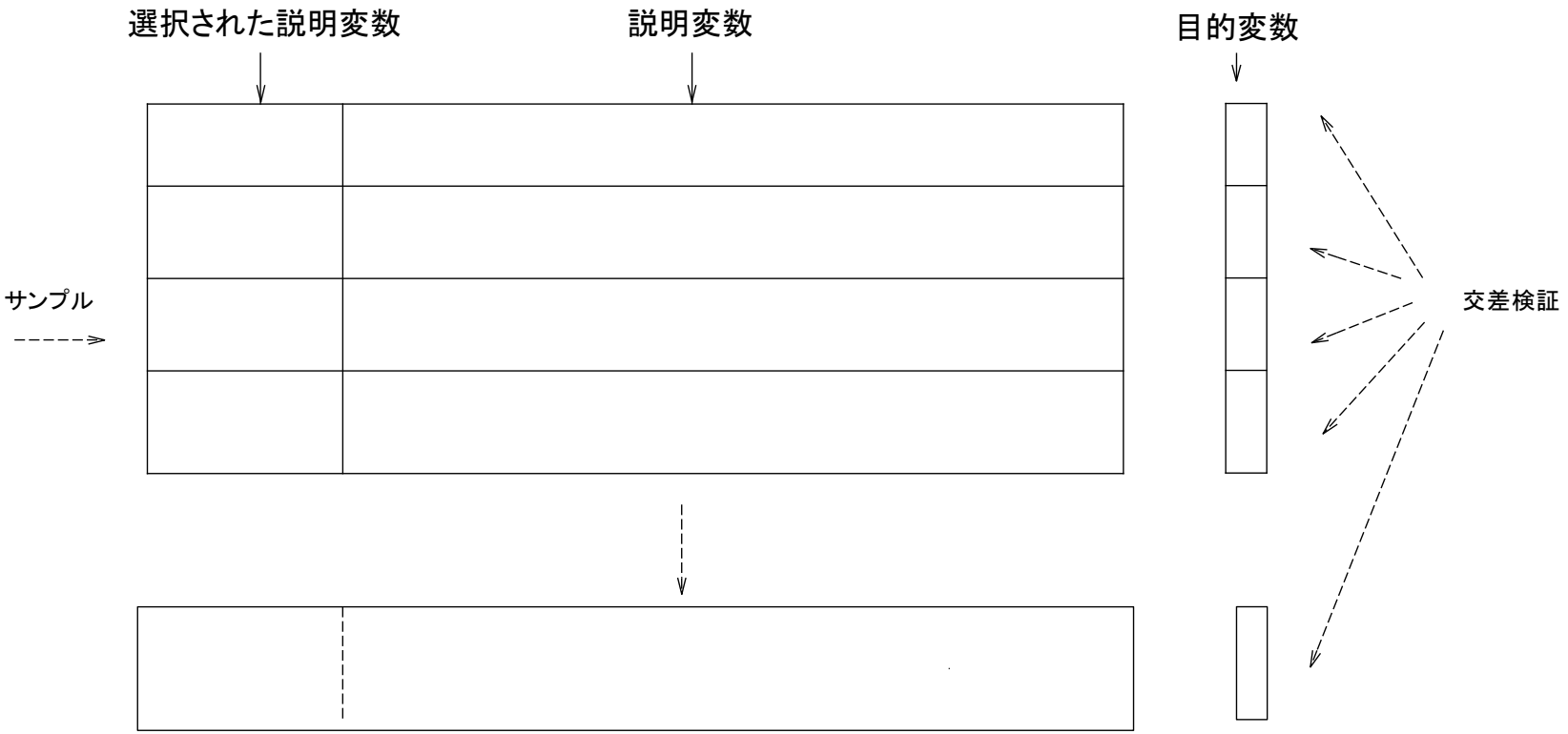
交差検証：正しい方法と誤った方法

- 誤った方法：交差検証をステップ2で適用する
- 正しい方法：交差検証をステップ1で適用する

誤った方法



正しい方法



ブートストラップ

- ブートストラップは、与えられた推定量や統計的学習方法に関連する不確実性を数量化するために使用できる、柔軟で強力な統計ツールである。
- たとえば、係数の標準誤差の推定値や、その係数の信頼区間を推定する際、ブートストラップを使うことができる。

「ブートストラップ」の由来

- ブートストラップは、*to pull oneself up by one's bootstraps*（自分でブートストラップ（長靴のつまみ革）を引っ張って自分を持ち上げる）に由来する。
- 18世紀の小説 "The Surprising Adventures of Baron Munchausen" by Rudolph Erich Raspe:

  > The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps. （Baronさんが湖の底に落ちて、自分の長靴のつまみ革をつかんで、自分を引き上げた）

- コンピュータサイエンスでの「ブートストラップ」という用語は、コンピュータを一連のコア命令から起動することを意味するものであり、語源は似ているが、今回の「bootstrap」とは異なる。

単純な例

- XとYがランダム変数である2つの金融資産に固定金額を投資することを考える。
- 手持ちの金額のうち割合αをXに投資し、残りの1−αをYに投資することにする。
- 投資の合計リスクまたは分散を最小化するαを選択することを望む。つまり、Var(αX + (1−α)Y)を最小化したいと考えている。

単純な例

- XとYがランダム変数である2つの金融資産に固定金額を投資することを考える。
- 手持ちの金額のうち割合αをXに投資し、残りの1−αをYに投資することにする。
- 投資の合計リスクまたは分散を最小化するαを選択することを望む。つまり、Var(αX + (1−α)Y)を最小化したいと考えている。
- 以下の式を用いてリスクを最小にすることができる

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

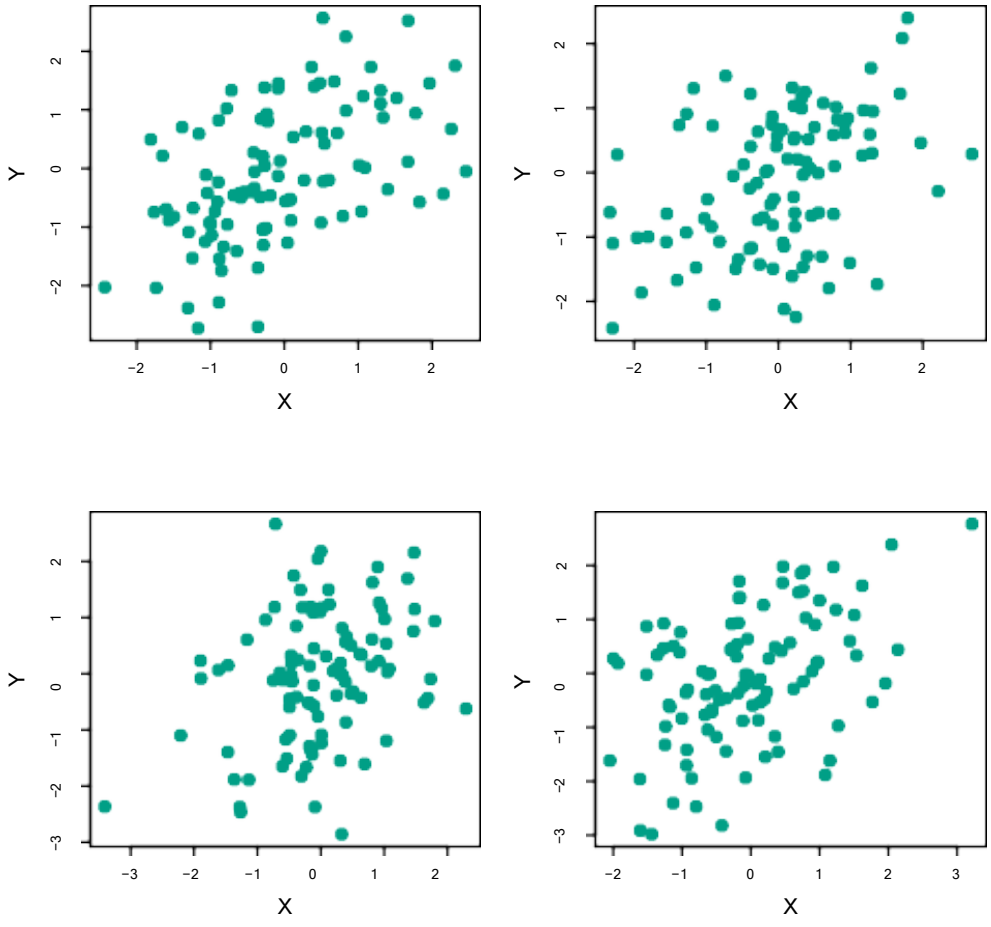
ここに$\sigma_X^2 = Var(X), \sigma_Y^2 = Var(Y), \sigma_{XY} = Cov(X, Y)$である。

単純な例

- しかし、$\sigma_X^2, \sigma_Y^2, \sigma_{XY}$は未知である。
- これらの$\hat{\sigma}_X^2, \hat{\sigma}_Y^2, \hat{\sigma}_{XY}$を過去のXとYの観測データより計算することができる。
- そして以下の式を使い、投資の分散を最小化するαを推定することができる。

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

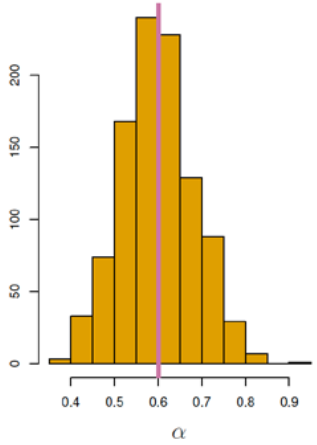
単純な例



各グラフには、シミュレーションによって得られた投資XとYの100組の結果が表示される。左から右に、上から下に、αの推定値は、0.576、0.532、0.657、0.651である。

単純な例

- $\hat{\alpha}$の標準誤差を推定するため、シミュレーションによってXとYを100組生成し、αを推定するプロセスを1000回繰り返した。
- これによりαの推定値を1000個得たことになる。これらを$\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}$と呼ぶことにする。
- スライド29ページの左の図はこの推定値のヒストグラムである。
- データを生成するときのパラメータは$\sigma_X^2 = 1, \sigma_Y^2 = 1.25, \sigma_{XY} = 0.5$としたので、真のαは0.6であることは既知である。



単純な例

- 1000個のαの推定値の平均は

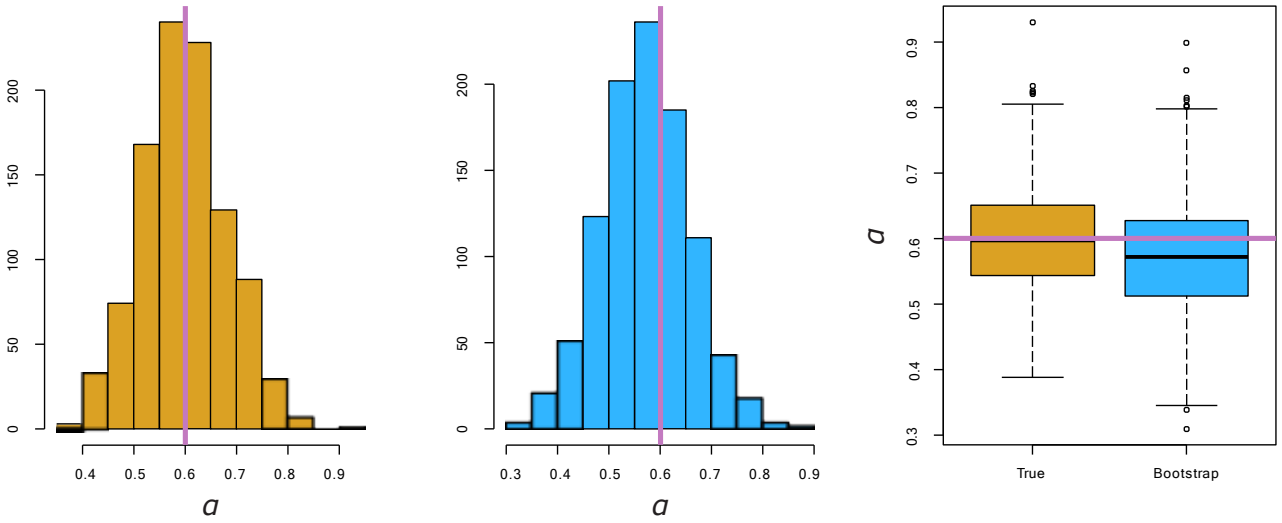
$$\bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000}\hat{\alpha}_r = 0.5996$$

- であり、真の値α＝0.6に非常に近い。また推定値の標準誤差は

$$\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$$

- これは$\hat{\alpha}$の精度に非常に近い：$SE(\hat{\alpha}_r)$≈0.083。
- 大まかに言えば母集団から無作為抽出した標本では、αと$\hat{\alpha}$の違いはおよそ平均0.08だといえる。

結果

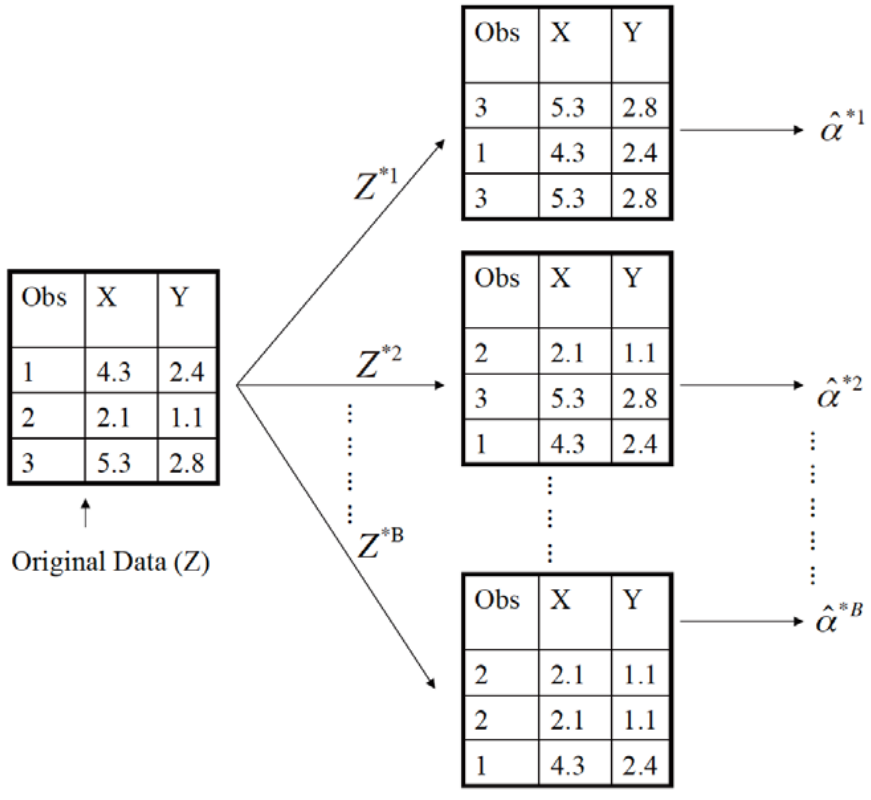


左：真の分布で1000個のデータをシミュレーションによって発生させ、αの推定値を得たときのヒストグラム。
中央：単一のデータセットから1000個のブートストラップ標本を取り出し、αの推定値を得たときのヒストグラム。
右：左と中央のグラフにおけるαの推定値の箱ひげ図、それぞれの図でピンク線は真のαの値を示す。

リアル世界に戻ろう

- 実際には、真の分布を知ることができないため、先程の手順では新たなデータを生成することはできない。
- しかし、ブートストラップを使うことで、実際に新たな標本を生成することなく$\hat{\alpha}$のばらつきを推定することができる。
- ブートストラップは、もともと手元にあるデータそのものから標本を抽出する方法であり、真の分布から独立したデータセットを繰り返し得る必要はない。
- ブートストラップデータセットは、もとのデータセットから同じサイズのデータを無作為に重複を許して抽出することによって作成される。つまりブートストラップ標本の中には全く同じデータが複数存在しても良いということである。

3個のデータだけを含むデータセット



3つのデータを含むデータセットでブートストラップを適用した場合の図解である。各ブートストラップ標本は、もとのデータから重複を許して抽出された3つのデータである。各ブートストラップ標本を用いてαの推定値を得ることができる。

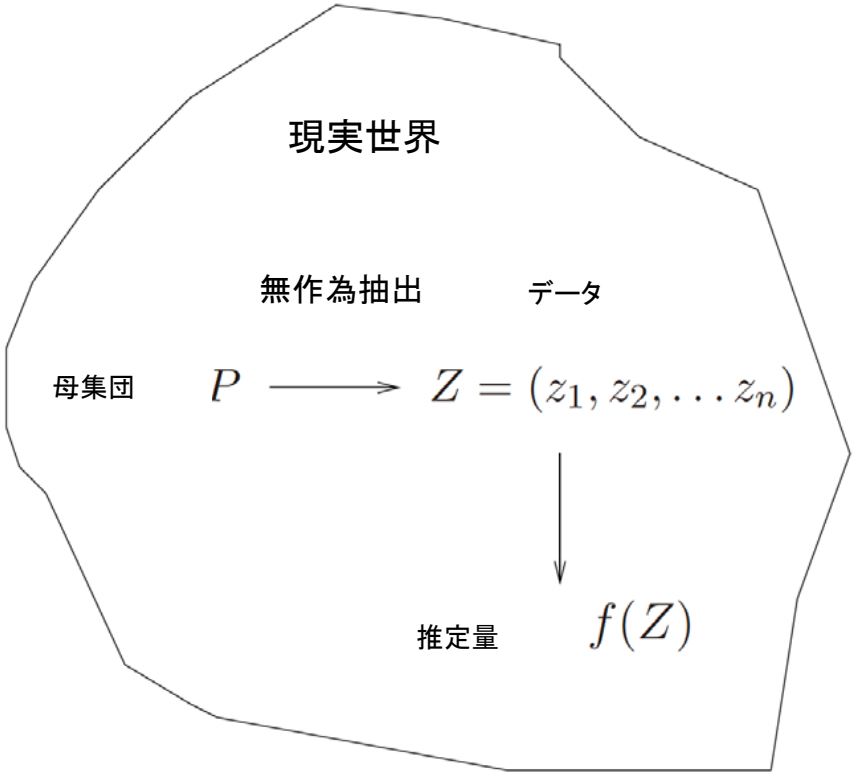
ブートストラップ

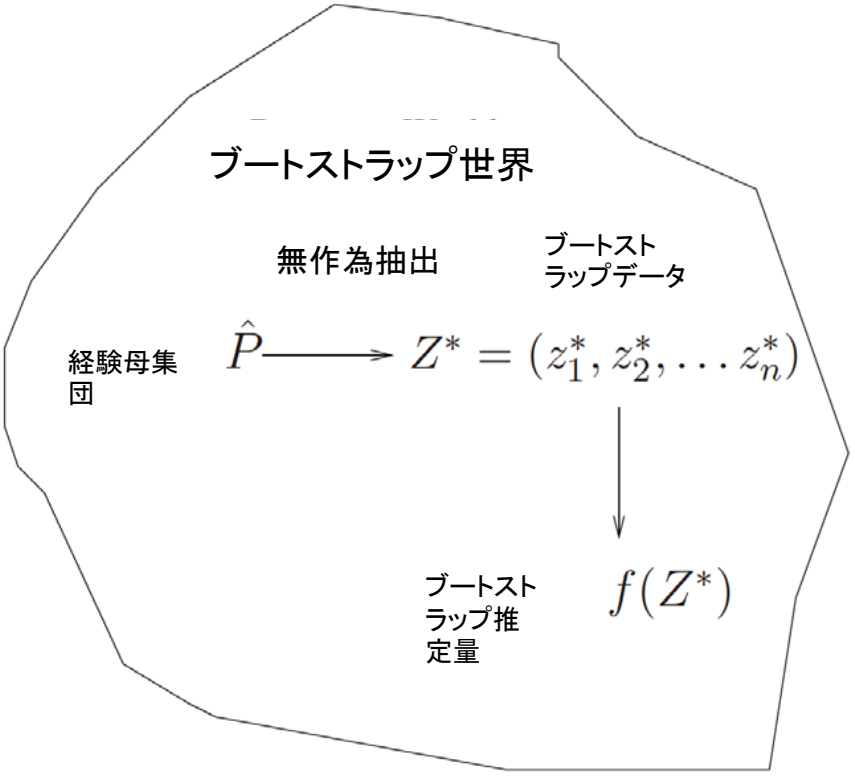
- 最初に作成されたブートストラップデータセットをZ^{*1}とする。Z^{*1}を使って推定したαを$\hat{\alpha}^{*1}$とする。
- 以上をB回（100または1000）繰り返し、B個の異なるブートストラップデータセット$Z^{*1}, Z^{*2}, \ldots, Z^{*B}$と、対応するαの推定値をB個$\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$を得る。
- 以下の式により、ブートストラップ推定値の標準誤差を求めることができる。

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2}$$

- これをもって元のデータセットから推定した$\hat{\alpha}$の標準誤差の推定値とする。

ブートストラップの世界観





ブートストラップの概要

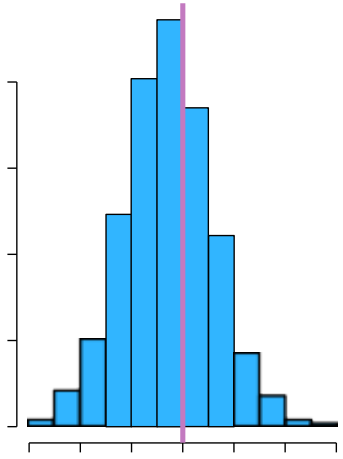
- より複雑なデータの場合、ブートストラップサンプルを生成する適切な方法を考えることが必要である。
- 例えば、データが時系列の場合、復元抽出を行ってはいけない（なぜ？）。

ブートストラップの概要

- より複雑なデータの場合、ブートストラップサンプルを生成する適切な方法を考えることが必要である。
- 例えば、データが時系列の場合、復元抽出を行ってはいけない（なぜ？）。
- 代わりに、連続する観測値のブロックを作成し、それらを復元抽出することができる。その後、抽出されたブロックを合わせて、ブートストラップデータセットを得ることができる。

ブートストラップのその他の使い道

- 主に推定量の標準誤差を求めるために使用される。
- また、母数の信頼区間を近似的に求めることができる。たとえば、下図のヒストグラムを見ると、1000個の値の5%と95%の分位点は(.43, .72)である。
- これは、真のαの90%信頼区間を近似的に表したものである。



ブートストラップのその他の使い道

- 主に推定量の標準誤差を求めるために使用される。
- また、母数の信頼区間を近似的に求めることができる。たとえば、下図のヒストグラムを見ると、1000個の値の5%と95%の分位点は(.43, .72)である。
- これは、真のαの90%信頼区間を近似的に表したものである。この信頼区間をどう解釈するか？

ブートストラップのその他の使い道

- 主に推定量の標準誤差を求めるために使用される。
- また、母数の信頼区間を近似的に求めることができる。たとえば、下図のヒストグラムを見ると、1000個の値の5%と95%の分位点は(.43, .72)である。
- これは、真のαの90%信頼区間を近似的に表したものである。この信頼区間をどう解釈するか？
- 上記の区間は、ブートストラップパーセンタイル信頼区間と呼ばれる。これは、ブートストラップから信頼区間を得るための多くの手法の中で、最も簡単な方法である。

ブートストラップは予測誤差を推定できるか？

- 交差検証では、K個の検証データセットのそれぞれは他のK-1個の検証データセットと重複することなく学習に使用される。これはその成功にとって重要である。

ブートストラップは予測誤差を推定できるか？

- 交差検証では、K個の検証データセットのそれぞれは他のK-1個の検証データセットと重複することなく学習に使用される。これはその成功にとって重要である。なぜ？

ブートストラップは予測誤差を推定できるか？

- 交差検証では、K個の検証データセットのそれぞれは他のK-1個の検証データセットと重複することなく学習に使用される。これはその成功にとって重要である。なぜ？
- ブートストラップを使用して予測誤差を推定するには、各ブートストラップ標本を訓練データとして使用し、元のデータセットを検証データとして使用することができる。
- ただし、各ブートストラップ標本には元のデータと重複する部分がかなりある。元のデータセットの約2/3が各ブートストラップ標本に表示される。

ブートストラップは予測誤差を推定できるか？

- 交差検証では、K個の検証データセットのそれぞれは他のK-1個の検証データセットと重複することなく学習に使用される。これはその成功にとって重要である。なぜ？
- ブートストラップを使用して予測誤差を推定するには、各ブートストラップ標本を訓練データとして使用し、元のデータセットを検証データとして使用することができる。
- ただし、各ブートストラップ標本には元のデータと重複する部分がかなりある。元のデータセットの約2/3が各ブートストラップ標本に表示される。どうやって証明できる？

ブートストラップは予測誤差を推定できるか？

- 交差検証では、K個の検証データセットのそれぞれは他のK-1個の検証データセットと重複することなく学習に使用される。これはその成功にとって重要である。なぜ？
- ブートストラップを使用して予測誤差を推定するには、各ブートストラップ標本を訓練データとして使用し、元のデータセットを検証データとして使用することができる。
- ただし、各ブートストラップ標本には元のデータと重複する部分がかなりある。元のデータセットの約2/3が各ブートストラップ標本に表示される。どうやって証明できる？
- このままでは、ブートストラップは真の予測誤差を大幅に過小評価してしまう。

ブートストラップは予測誤差を推定できるか？

- 交差検証では、K個の検証データセットのそれぞれは他のK-1個の検証データセットと重複することなく学習に使用される。これはその成功にとって重要である。なぜ？
- ブートストラップを使用して予測誤差を推定するには、各ブートストラップ標本を訓練データとして使用し、元のデータセットを検証データとして使用することができる。
- ただし、各ブートストラップ標本には元のデータと重複する部分がかなりある。元のデータセットの約2/3が各ブートストラップ標本に表示される。どうやって証明できる？
- このままでは、ブートストラップは真の予測誤差を大幅に過小評価してしまう。なぜ？
- 元のデータセットを訓練データ、ブートストラップ標本を検証データとして使う場合、結果は悪くなる。

オーバーラップを解消する

- 問題を部分的に解決するには、現在のブートストラップ標本で偶然発生しなかった観測値に対してのみ予測を使用することができる。
- しかし、この方法は複雑になり、最終的には交差検証がよりシンプルで魅力的なアプローチとなる。これにより、予測誤差を推定することができる。

プレバリデーション

- マイクロアレイや他のゲノム研究において、大量の「バイオマーカー」から派生した疾患の結果を予測する予測変数を、標準的な臨床予測変数と比較することは重要な課題である。
- バイオマーカー予測変数を派生させたデータセットと同じデータセットで比較すると、バイオマーカー予測変数に偏った結果となることがある。
- プレバリデーションは、2つの予測変数のセット間でより公平な比較をするために使用される。

例

以下は、van't Veerらの論文（Nature、2002）において問題が発生した例である。彼らのマイクロアレイデータは、乳癌の研究から取得され、78の症例に渡って4918の遺伝子が測定されている。良好な予後グループには44の症例があり、悪い予後グループには34の症例がある。次のようにして「マイクロアレイ」予測変数が構築された。

1. 78のクラスラベルと最も相関の高い70の遺伝子が選択された。
2. これらの70の遺伝子を使用して、最近傍重心分類器C(x)が構築された。
3. 78のマイクロアレイに分類器を適用すると、各症例iに対してジコトモス予測変数zi = C(xi)が得られる。

結果

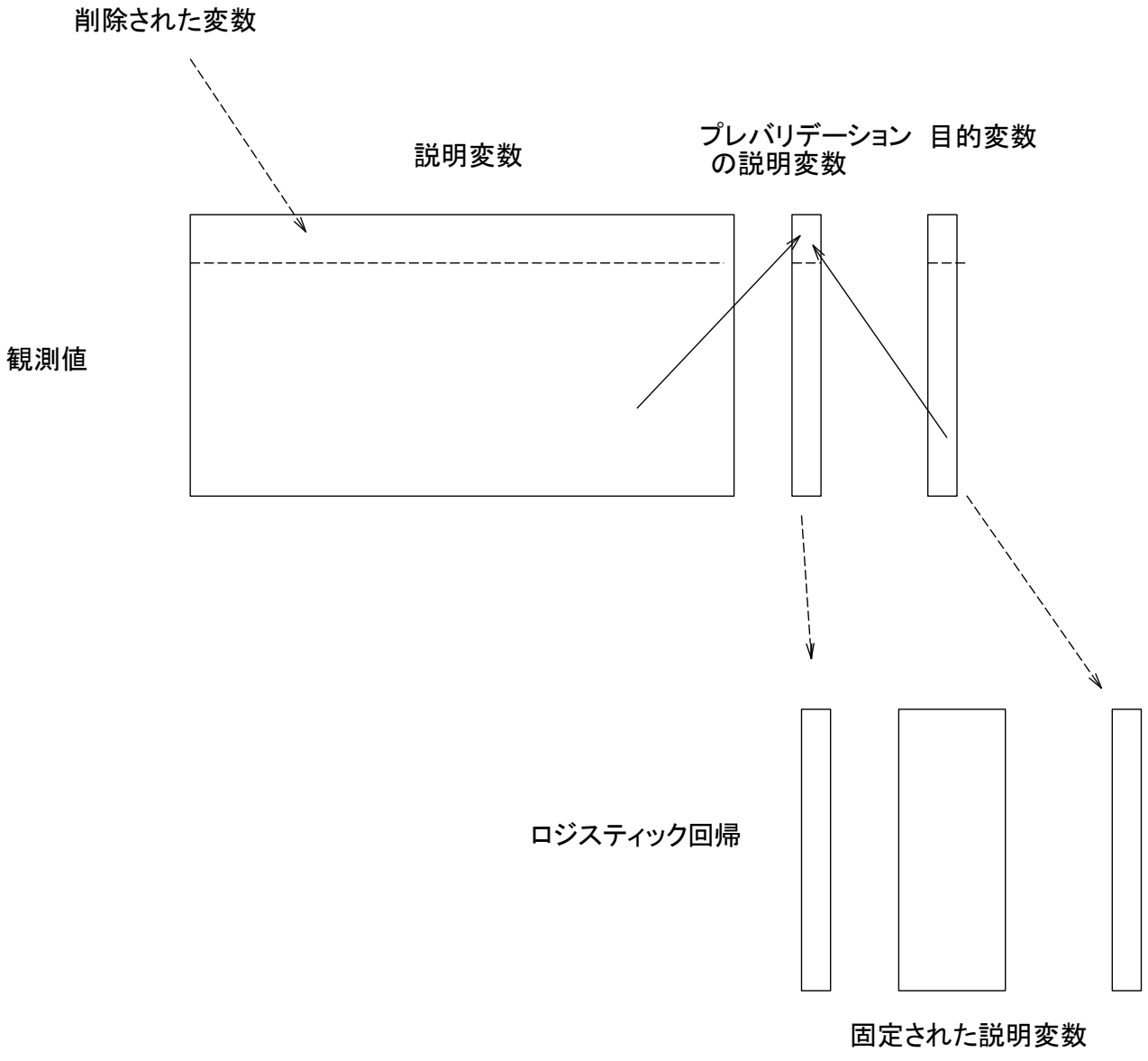
目的変数prognosisを用いたロジスティック回帰を用いて、マイクロアレイ予測変数といくつかの臨床予測変数を比較する。

Model	Coef	Stand. Err.	Z score	p-value
		Re-use		
microarray	4.096	1.092	3.753	0.000
angio	1.208	0.816	1.482	0.069
er	-0.554	1.044	-0.530	0.298
grade	-0.697	1.003	-0.695	0.243
pr	1.214	1.057	1.149	0.125
age	-1.593	0.911	-1.748	0.040
size	1.483	0.732	2.026	0.021
		Pre-validated		
microarray	1.549	0.675	2.296	0.011
angio	1.589	0.682	2.329	0.010
er	-0.617	0.894	-0.690	0.245
grade	0.719	0.720	0.999	0.159
pr	0.537	0.863	0.622	0.267
age	-1.471	0.701	-2.099	0.018
size	0.998	0.594	1.681	0.046

プレバリデーションの考え方

- 適応的に導出された予測変数と、事前定義された予測変数の比較のために設計された。
- プレバリデーションの考え方は、適応的な予測変数の「事前検証」を構成することである。特に、目的変数yを「見ていない」より「公正な」手法である。

プレバリデーションの手順



プレバリデーションの詳細（この例の場合）

1. 6つのケースを含む同じサイズ（K = 13）のグループに分割する。
2. 1つのグループを取り置く。残りの12グループのデータのみを使用して、クラスラベルとの絶対相関が少なくとも.3以上の特徴量を選択し、最近傍中心分類ルールを構成する。
3. 構成したルールを使用して、第13グループのクラスラベルを予測する。
4. 各13グループについてステップ2と3を実行し、78のケースそれぞれに対してプレバリデーションのマイクロアレイ予測変数$\tilde{z}_i$を生成する。
5. プレバリデーションによるマイクロアレイ予測変数と6つの臨床予測変数を用いてロジスティック回帰モデルをあてはめる。

ブートストラップとパーミュテーション検定の比較

- ブートストラップは推定された母集団からサンプリングし、その結果を使用して標準誤差と信頼区間を推定する。
- パーミュテーション検定は、データの推定されたヌル分布からサンプリングし、これを使用して仮説検定のP値と誤検出率を推定する。
- ブートストラップは、単純な状況で帰無仮説を検定するために使用できる。例えば、θ=0が帰無仮説である場合、θの信頼区間がゼロを含むかどうかを確認する。
- ブートストラップを帰無分布からサンプリングするように適応することもできる（EfronとTibshiraniの書籍「An Introduction to the Bootstrap」（1993）、第16章を参照）。ただし、パーミュテーション検定に比べて実質的な利点はない。

第6章：線形モデル選択と正則化
-Linear Model Selection and Regularization-

- 導入
- 部分集合選択
- 変数増加法
- 変数減少法
- テスト誤差の推定
- 検証と交差検証
- 縮小法とリッジ回帰
- Lasso
- チューニングパラメータ選択
- 次元削減法
- 主成分回帰と部分最小二乗

第6章：線形モデル選択と正則化
-Linear Model Selection and Regularization-

- 線形モデルは次のよう

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- 本講義では、いくつかの方法で線形モデルの枠組みを拡張する. テキストの7章をカバーする講義では、非線形性を取り入れるために線形モデルを一般化する. ただし加法性は保ったままとする.
- 8章をカバーする講義では、より一般的な非線形モデルを扱う.

線形モデルの活用

- 線形モデルはその単純さにもかかわらず、解釈性の観点から明確な利点があり、時折良い予測性能を発揮する.
- そこで、今回の講義では単純な線形モデルに対し、最小二乗法ではない他のいくつかの推定方法を用いる事で改善がなされ得ることを議論する.

なぜ最小二乗法以外のものを用いるか？

- 予測精度: 特に$p > n$のとき、分散をコントロールするため.
- モデルの解釈性: 無関係な特徴を除く事、つまり、対応する係数の推定値を0とする事、によってより解釈がしやすいモデルを得ることができる. この後、自動的に特徴選択を行ういくつかの方法を提示する.

3種類の方法

- 部分集合選択. p個の内、応答に関連すると思われる予測変数の部分集合を特定する. 次に減らされた変数を用いて最小二乗によるモデルのあてはめを行う.
- 縮小. 全ての p 個の予測変数を用いてモデルをあてはめるが、さ推定された係数が最小二乗推定と比べると0に縮小される. この縮小は(正則化としても知られ)分散を減らす効果を持ち、変数選択を行う事も出来る.
- 次元削減. p個の予測変数をM-次元の部分空間に射影する、ただし$M < p$. これはM個の異なる、変数の線形結合や射影を計算する事によって行われる. これらのM個の射影は、最小二乗によって線形回帰モデルのあてはめのための予測変数として用いられる.

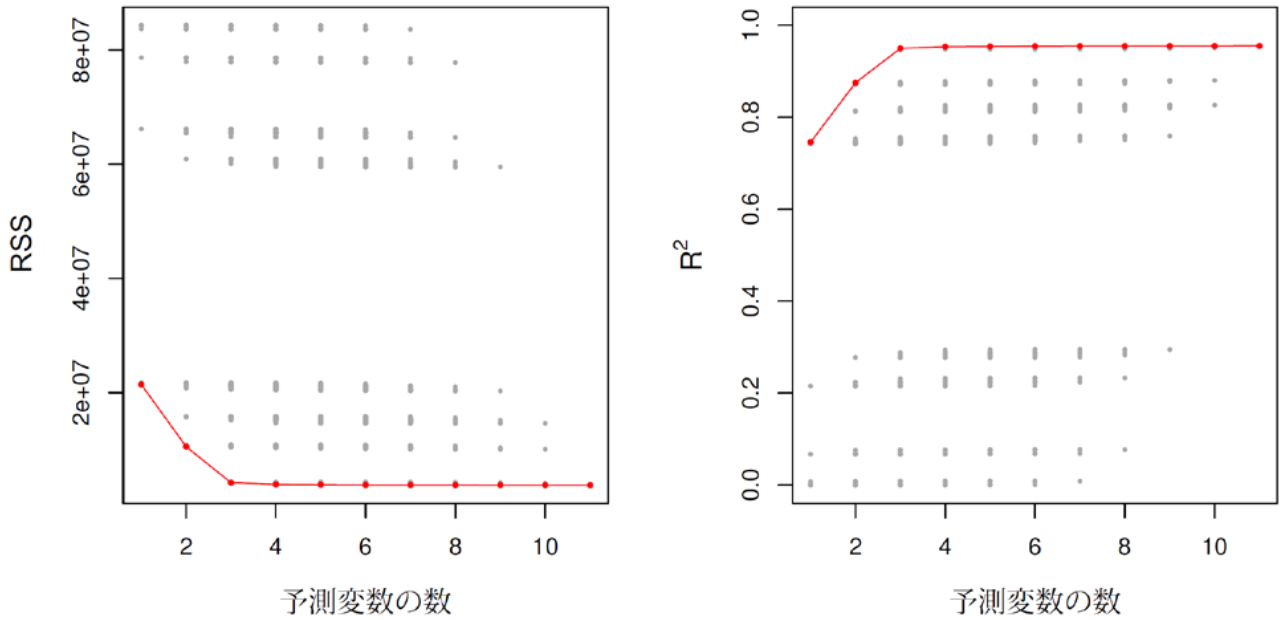
部分集合選択

最良部分集合とステップワイズなモデル選択の方法

最良の部分集合選択

1. $\mathcal{M}_0$によって予測変数を1つも含まないnull モデルを表す. このモデルは単に各観測の標本平均により予測を行う.
2. $k=1, 2, \quad p$に対し
   (a) ちょうどk個の予測変数を含む$\binom{p}{k}$個のモデル全てをあてはめる.
   (b) この$\binom{p}{k}$個のモデルの内、最良なものを選び、$\mathcal{M}_k$とする. ここで最良とは最小のRSS、つまり最大のR^2を持つものとする.
3. 交差検証予測誤差、C_p (AIC)、BICや調整済みR^2を用いて$\mathcal{M}_0, \ldots, \mathcal{M}_p$の内、最良のモデルをただ1つ選ぶ.

例- クレジットデータセット



クレジットデータセットの10個の予測変数の部分集合を含む各モデルに対して、RSSとR^2を図示している. 赤色で示された境界が、与えられた予測変数の数の下で、RSSやR^2に基づいた最良なモデルを表している. データセットは10個しか予測変数を含まないが、x軸が1から11まであるのは、変数の内の1つがカテゴリカルで3つの値を取り得て2つのダミー変数を作っているからである.

他のモデルへの拡張

- ここまで最小二乗回帰に対する最良部分集合選択を紹介したが、同じ考え方を他のモデル、例えばロジスティック回帰にも適用する事ができる.
- 偏差-最大対数尤度のマイナス2倍-がモデルの広いクラスに対するRSSの役割を果たす.

ステップワイズ選択

- 計算機的理由として、とても大きなpに対しては最良部分集合選択を適用する事は出来ない. 何故か？
- 最良部分集合選択はpが大きい時に統計的問題にも苦しめられる. 探索空間が大きいほど、訓練データに対して良さそうなモデルを見つけられる可能性は高くなるが、将来得られるデータに対する予測性能はさほどかもしれない.
- こうして、大きな探索空間は過適合につながり、係数の推定値の分散を大きくする.
- これらの理由から、モデルのより制約された集合から探索を行うステップワイズな方法は、最良部分集合選択に代わる魅力的な方法となる.

変数増加法

- 変数増加法(前向きステップワイズ選択)は、1つも予測変数を含まないモデルから始まり、モデルに予測変数を加えていく. モデルに全ての予測変数が含まれるまで行う.
- 特に、各ステップでは、あてはまりに関して最大の追加の改良を行うような変数が、モデルに追加される.

詳細

変数増加法(前向きステップワイズ選択)

1. $\mathcal{M}_0$によって予測変数を1つも含まないnull モデルを表す.
2. $k = 0,1, \ldots, p-1$に対し
   2.1. 1つ追加の予測変数を$\mathcal{M}_k$に加えたp $- k$個のモデルを考える.
   2.2. これらの$p - k$個の内、最良なものを選び、$\mathcal{M}_{k+1}$と呼ぶ. ここで最良とは最小のRSS、最大のR^2を持つものとする.
3. 交差検証予測誤差、C_p (AIC)、BICや調整済みR^2を用いて$\mathcal{M}_0, \ldots, \mathcal{M}_p$の内、最も良いモデルをただ1つ選ぶ.

変数増加法について更に

- 最良部分集合選択に対する計算機上の利点は明確.
- p個の予測変数の部分集合を含むような2^p個の全てのモデルから最良なモデルを見つける事を保証はしない. 何故か？
  例を与えよ.

クレジットデータの例

変数の数	最良部分集合	変数増加法
1	rating	rating
2	rating, income	rating, income
3	rating, income, student	rating, income, student
4	cards, income student, limit	rating, income, student, limit

クレジットデータセットに対して、最良部分集合選択と
変数増加法による、初めの4つの選択されたモデル.
初めの3つは一致しているが、4番目は異なっている.

変数減少法

- 変数増加法のように、変数減少法(後ろ向きステップワイズ選択)は最良部分集合選択に代わる効率的な方法を与える.
- しかし、変数増加法と違い、全てのp個の予測変数を含む最小二乗から始まる. 続いて、繰り返し、最も有用でない予測変数を除いていく.

変数減少法 : 詳細

変数減少法(後ろ向きステップワイズ選択)

1. $\mathcal{M}_p$によってp個の全ての予測変数を含むfull モデルを表す.
2. $k = p, p-1, \ldots, 1$に対し
   - 2.1. $\mathcal{M}_k$ から1つ予測変数だけを除いたk個のモデルを考える.
   - 2.2. これらのk個の内、最良なものを選び、$\mathcal{M}_{k-1}$と呼ぶ. ここで最良とは最小のRSS、最大のR^2を持つものとする.
3. 交差検証予測誤差、C_p (AIC)、BICや調整済みR^2を用いて$\mathcal{M}_0, \ldots, \mathcal{M}_p$の内、最も良いモデルをただ1つ選ぶ.

変数減少法について更に

- 変数増加法のように、変数減少法は$1 + p(p + 1)/2$個のモデルだけから探索を行う. 最良部分集合選択を適用するにはpが大きすぎるような場合でも適用が可能である.
- 変数増加法と同様、減数減少法はp個の予測変数の部分集合を含む最良なモデルを導く事は保証しない.
- 変数減少法はサンプルサイズnが変数の数pより大きい事が必要となる. (これはfullモデルのあてはめが可能であるため.) 対して、変数増加法は$n < p$の時でも使え、pがとても大きい時唯一実行可能な部分集合法である.

最適なモデルの選択

- 全ての予測変数を含むモデルが最小のRSSと最大のR^2を持つ. これらの値が訓練誤差に関連するためである.
- 訓練誤差が低いモデルではなく、テスト誤差が低くなるモデルを選びたい. 訓練誤差はたいていテスト誤差の推定量としては精度が低い事に注意する.
- よって、RSSとR^2は予測変数の数が異なるモデルの中で最良なモデルを選択するのには適してはいない.

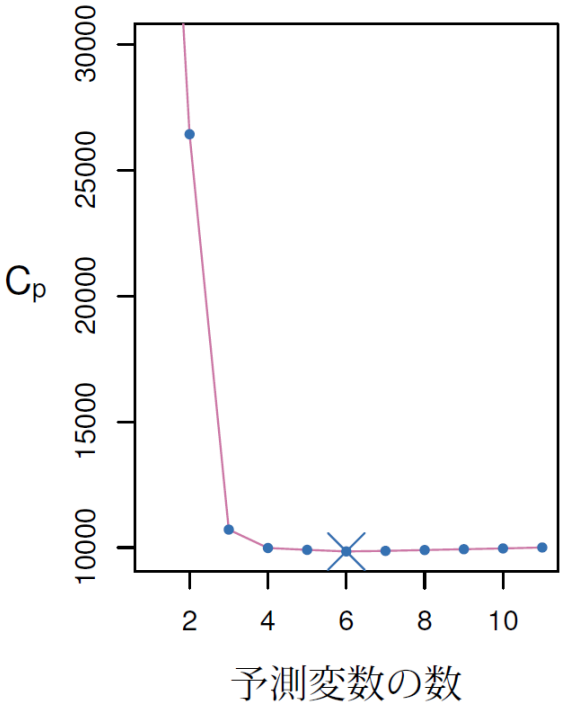
テスト誤差の推定: 2つの方法

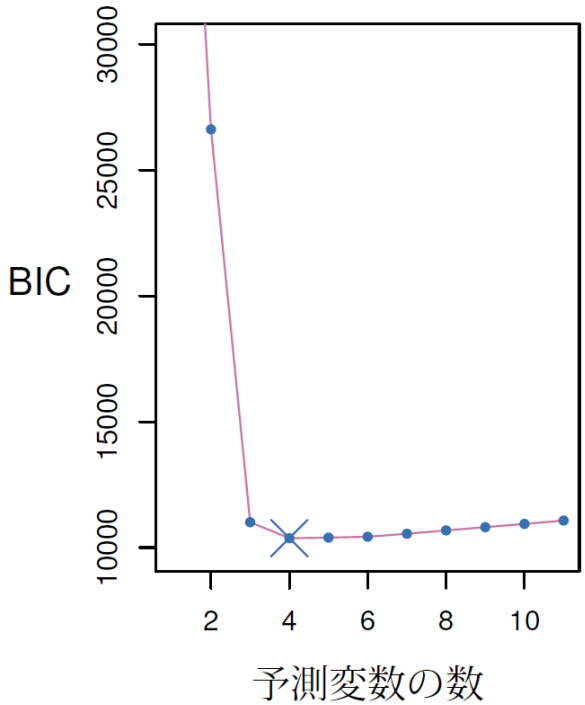
- 過適合によるバイアスを考慮して訓練誤差を調整する事で、間接的にテスト誤差を推定できる.
- 以前の講義で議論したように、検証セットを用いる方法や交差検証法を用いて、直接的にテスト誤差を推定できる.
- どちらの方法も次に説明する.

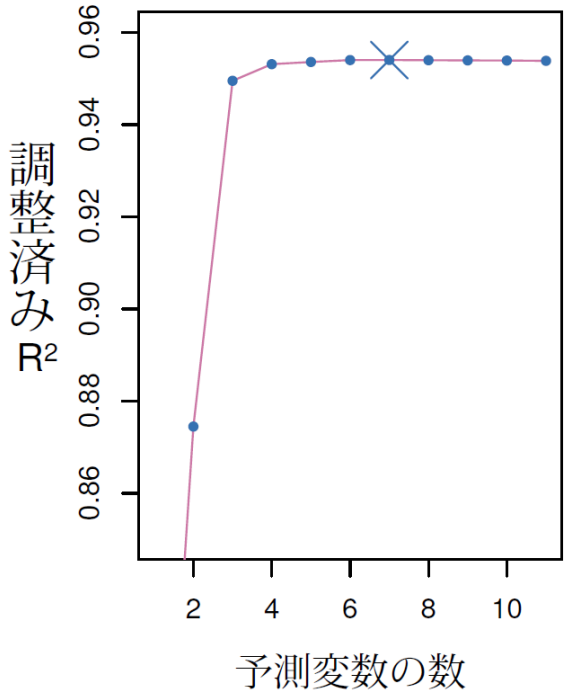
C_p, AIC, BICと調整済みR^2

- これらの方法はモデルサイズに関して訓練誤差を調整する. 変数の数が異なるモデルの中から選択を行うために用いる事が出来る.
- 次の図は、クレジットデータセットの最良部分集合によって生成される各サイズでの最良なモデルに関するC_p, BICと調整済みR^2を図示している.

クレジットデータの例


C_p
10000
15000
20000
25000
30000
2
4
6
8
10
予測変数の数


BIC
10000
15000
20000
25000
30000
2
4
6
8
10
予測変数の数


調整済み R^2
0.86
0.88
0.90
0.92
0.94
0.96
2
4
6
8
10
予測変数の数

いくつかの詳細

- マローのC_p:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

ただし、dはパラメータの総数で、$\hat{\sigma}^2$は各応答に伴う誤差ϵの分散の推定値.

- AIC基準は様々なクラスのモデルの最尤法によるあてはめに対して定義される

$$AIC = -2\log L + 2 \cdot d$$

ただしLは推定されたモデルに関する尤度関数の最大値

- ガウス誤差を持った線形モデルの場合、最大尤度や最小二乗は同じ事で、C_pとAICは等しい. これを示せ.

BICの詳細

$$BIC = \frac{1}{n}(RSS + \log(n) d\hat{\sigma}^2)$$

- C_pのように、BICはテスト誤差が小さいモデルに対して小さな値を取りやすいため、一般にBICが最小となるモデルを選択する.
- BICはC_pで$2d\hat{\sigma}^2$が用いられている部分を$\log(n) d\hat{\sigma}^2$で置き換える. ただし、nは観測数.
- $n > 7$に対して$\log n > 2$なので、BICは一般的に変数が多いモデルに対してより重い罰則を課す. よってC_pより小さなモデルの選択がなされる. スライド19の図を見よ.

調整済みR^2

- d個の変数を持った最小二乗モデルに対して、調整済みR^2は以下のように計算される

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}.$$

  ただしTSSは総二乗和.
- C_p, AICとBICが小さな値である事はテスト誤差の低いモデルである事を示すが、大きな値の調整済みR^2は小さなテスト誤差のモデルを示す.
- 調整済みR^2の最大化は$RSS/(n-d-1)$の最小化に等しい. RSSはモデルの変数の数が増えると常に減少するが、分母に含まれるdによって $RSS/(n-d-1)$は増加するかもしれないし減少するかもしれない.
- R^2と違って、調整済みR^2はモデルが不要な変数を含む事に対して代価を払うようになっている. スライド19の図を見よ.

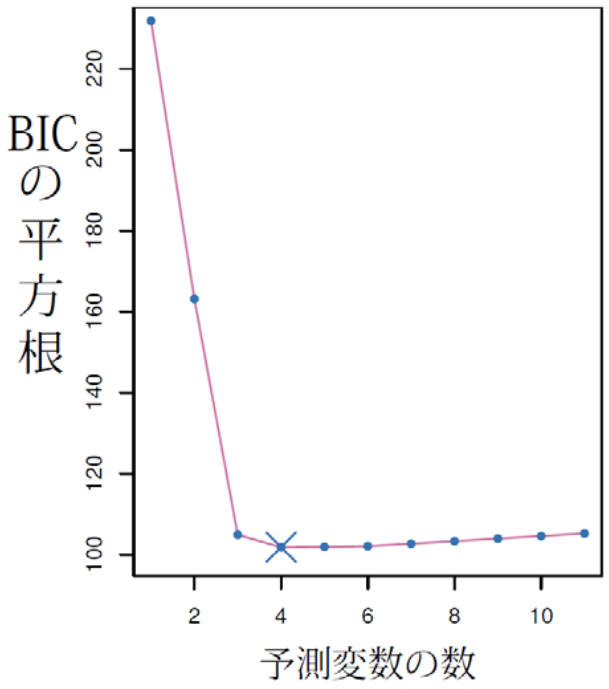
検証と交差検証

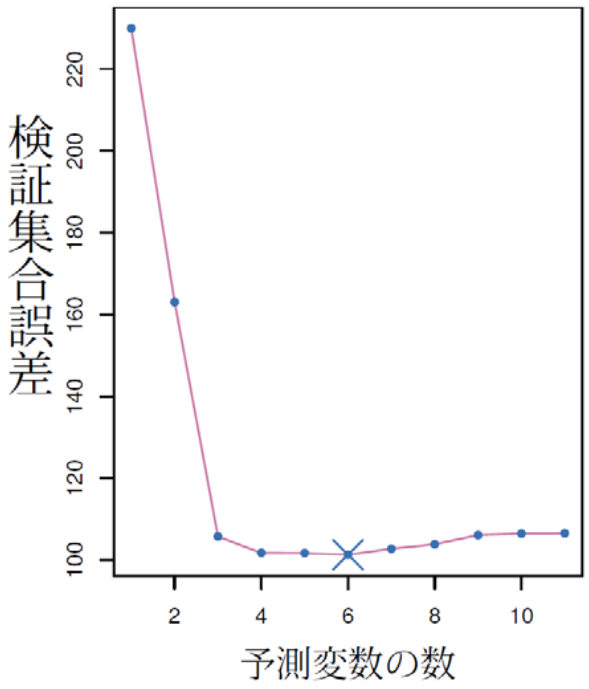
- モデルサイズ$k = 0,1,2,$ ...によってインデックスされた$\mathcal{M}_k$が各方法によって与えられる. ここでの課題は$\widehat{k}$の選択である. 選択がされると、$\mathcal{M}_{\widehat{k}}$が返される.

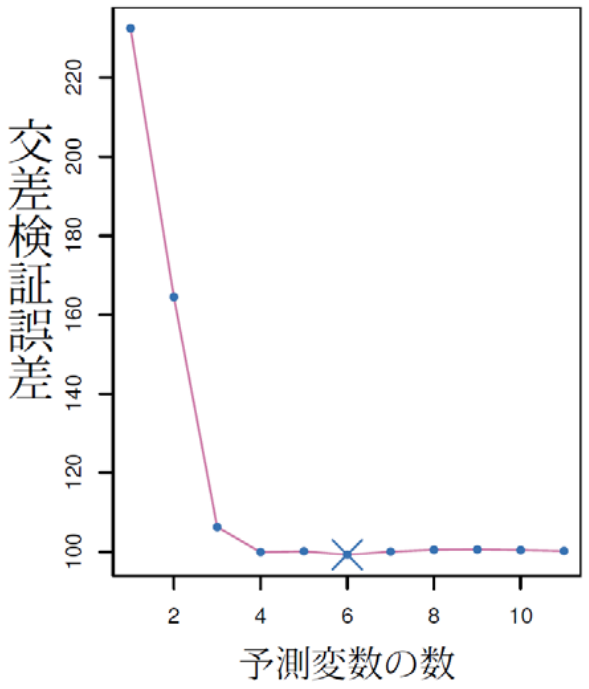
検証と交差検証

- モデルサイズ$k = 0,1,2,\ldots$によってインデックスされた$\mathcal{M}_k$が各方法によって与えられる. ここでの課題は$\widehat{k}$の選択である. 選択がされると、$\mathcal{M}_{\widehat{k}}$が返される.
- 各モデルに対して、検証集合誤差か交差検証誤差を計算する. 推定されたテスト誤差が最小となるようにkを選択する.
- この方法は、AIC, BIC, C_pや調整済みR^2に対して、テスト誤差の直接的な推定値を与えるという利点がある. また、誤差分散σ^2の推定値を必要としない.
- モデルの自由度の特定や誤差分散σ^2の推定が難しい時であっても、モデル選択に関する課題の広い範囲で用いられる.

クレジットデータの例


BICの平方根
100
120
140
160
180
200
220
2
4
6
8
10
予測変数の数


検証集合誤差
100
120
140
160
180
200
220
2
4
6
8
10
予測変数の数


交差検証誤差
100
120
140
160
180
200
220
2
4
6
8
10
予測変数の数

前の図の説明

- 訓練セットとしてランダムに観測の4分の3を選んで残りを検証セットとし、検証誤差を計算した.
- 交差検証誤差は$k = 10$分割を用いて計算した. この場合、検証や交差検証は共に6変数のモデルを導く.
- しかし、いずれの3つの方法も4,5,6変数のモデルはテスト誤差の観点からはおおよそ等しい.
- この設定では、1標準誤差ルールを用いてモデル選択が出来る. 初めに、各モデルサイズに対してテストMSEの推定値の標準誤差を計算する. さらに、テスト誤差の推定値が最小の点から1標準誤差の中に入るものの内、最小となるモデルを選ぶ. この根拠は何か？

縮小法

リッジ回帰とLasso

- 部分集合選択法は予測変数の部分集合を含む線形モデルをあてはめるために最小二乗を用いる.
- 代わりに、係数の推定値を制約や正則する方法、つまり係数の推定値を0に縮小する方法を用いてp個の予測変数全てを含むモデルをあてはめる.
- そのような制約によりあてはめの改良が何故なされるのかは直ちには明らかではない. しかし、係数の推定値の縮小が分散の減少に大きくつながることが分かる.

リッジ回帰

- 最小二乗法による推定では、次の値を最小化する $\beta_0, \beta_1, \ldots, \beta_p$を求めた.

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

- 対して、リッジ回帰による係数推定値$\hat{\beta}^R$は次を最小化する

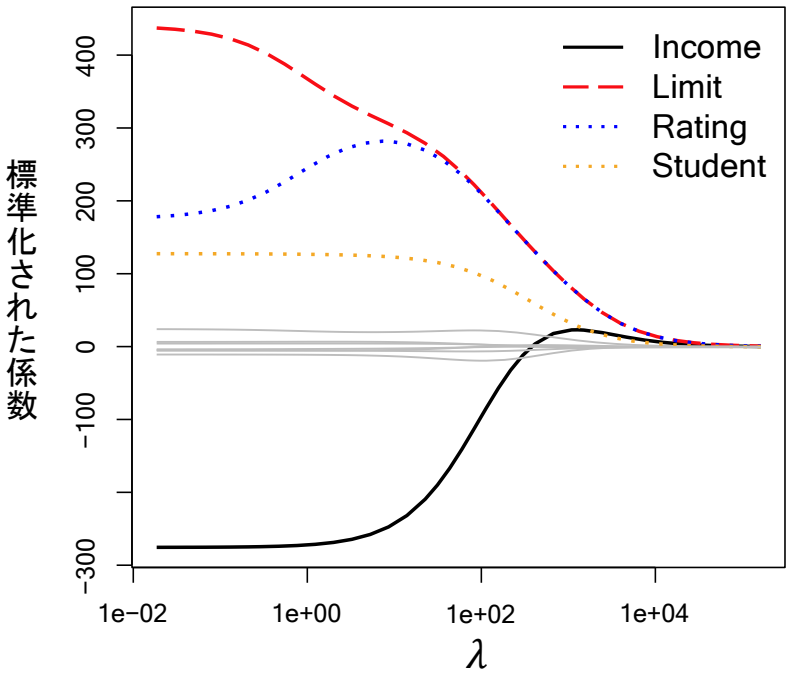
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

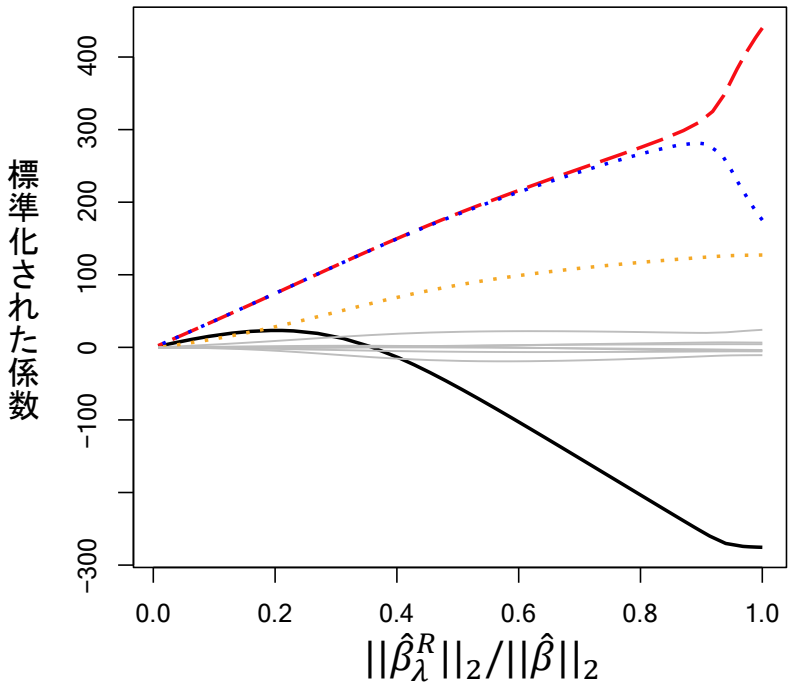
ただし$\lambda \geq 0$はチューニングパラメータで、それぞれの場合で決定される.

リッジ回帰: 続き

- 最小二乗のように、リッジ回帰はRSSを小さくする事によってデータに当てはまりの良い係数の推定値を求める方法である.
- しかし、第2項の$\lambda \sum_j \beta_j^2$は縮小罰則と呼ばれ、$\beta_1, \ldots, \beta_p$が0に近い時に小さな値をとる. β_jの推定値を0に縮小する効果を持つ.
- チューニングパラメータλは回帰係数の推定値に際して、これらの2つの項の相対的な重みをコントロールする.
- λの良い値を選択する事は重要で、交差検証はこのために用いられる.

クレジットデータの例


Income
Limit
Rating
Student
標準化された係数
400
300
200
100
0
−100
−300
1e−02
1e+00
1e+02
1e+04
λ


標準化された係数
400
300
200
100
0
−100
−300
0.0
0.2
0.4
0.6
0.8
1.0
$||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$

前の図の詳細

- 左図で、各曲線はλの関数として10個の変数のそれぞれのリッジ回帰の係数の推定値をプロットしている.
- 右図は左図と同じリッジ回帰の係数の推定値を図示しているが、x軸をλの代わりに$||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$としている. ただし$\hat{\beta}$は最小二乗による係数の推定値.
- $||\beta||_2$はベクトルのℓ_2ノルムを表している. $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$と定義される.

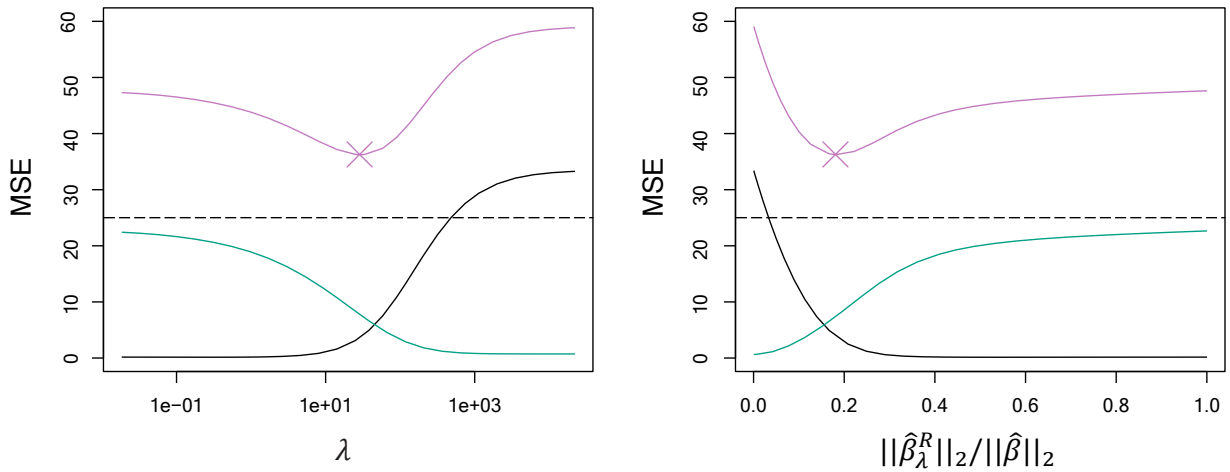
リッジ回帰: 予測変数のスケーリング

- 通常の最小二乗による係数推定値はスケール不変である: X_jをc倍すると最小二乗による係数推定値は$1/c$倍される. つまり、j番目の予測変数のスケールによらず$X_j\hat{\beta}_j$は変わらないままである.
- 対して、リッジ回帰の係数の推定値は、予測変数に定数をかけると大きく変化する. これはリッジ回帰の目的関数の罰則項の係数値の2乗和による.
- よって、次のような予測変数の標準化の後、リッジ回帰を適用する.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$

最小二乗に対してリッジ回帰を用いるべき理由は？

バイアスと分散のトレードオフ



$n = 50$個の観測と$p = 45$個の0でない係数を持つ予測変数からなるシミュレーションデータ. シミュレーションデータへのリッジ回帰予測に対する、2乗バイアス(黒)、分散(緑)、テスト平均二乗誤差(紫). λと$||\hat{\beta}_\lambda^R||_2/||\hat{\beta}||_2$の関数として表している. 点線は$MSE$のあり得る最小値を示している. 紫の×は$MSE$が最小となるリッジ回帰モデルを示している.

Lasso

- リッジ回帰は1つ明確な欠点がある. 変数の部分集合だけを含むようなモデルが選択される部分集合選択と違って、
  リッジ回帰による最終的なモデルは全てのp個の予測変数を含む.
- Lassoはこの欠点を克服した、リッジ回帰に代わる最近の手法である. Lassoによる係数推定値$\hat{\beta}_\lambda^L$は次を最小化する

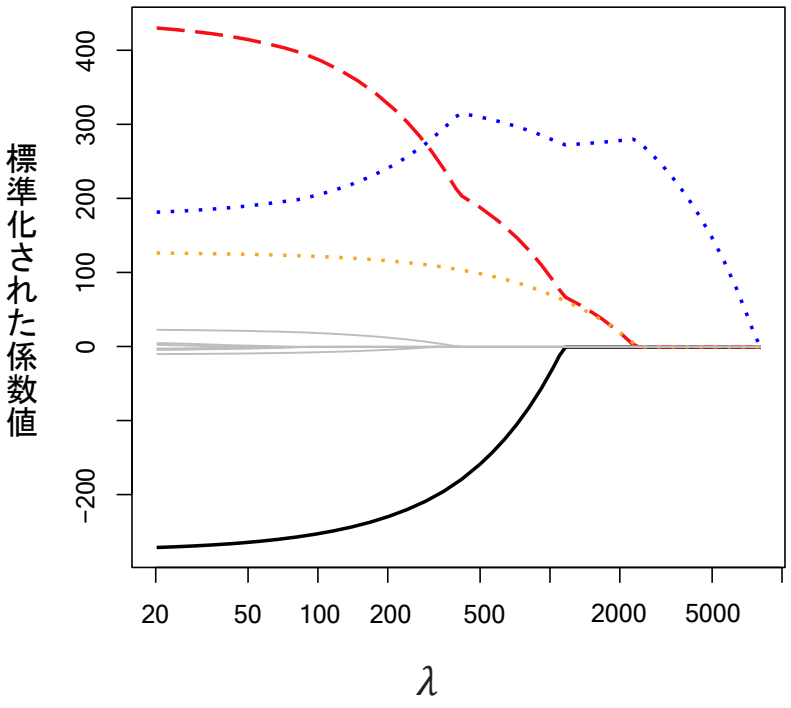
$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=RSS+\lambda\sum_{j=1}^{p}|\beta_j|$$

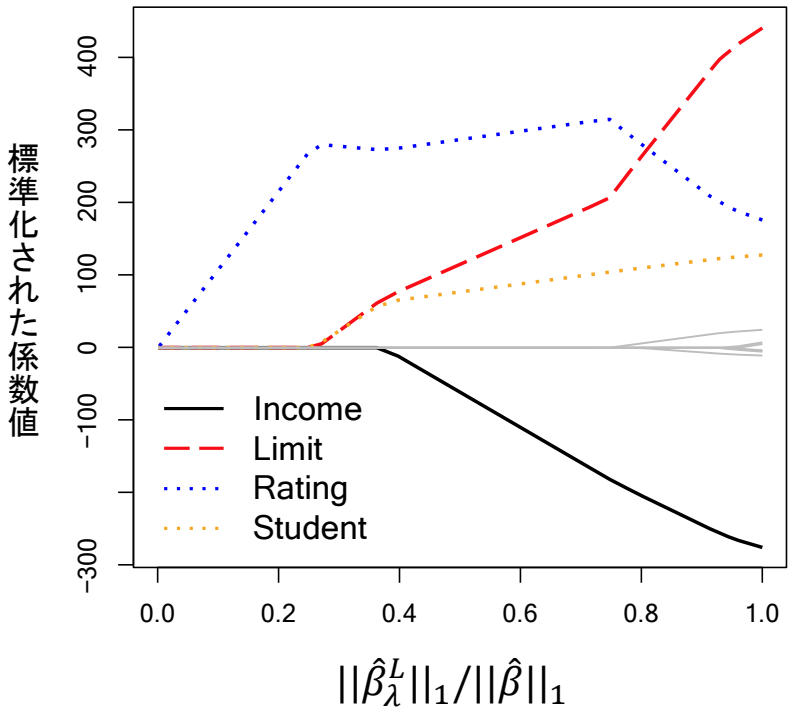
- 統計的な用語では、lassoはℓ_2罰則の代わりにℓ_1罰則を用いる.
  係数ベクトルβのℓ_1ノルムは$||\beta||_1=\sum|\beta_j|$によって与えられる.

Lasso: 続き

- リッジ回帰のように、lassoは係数の推定値を0に縮小させる.
- しかし、lassoの場合、ℓ_1罰則は、チューニングパラメータλが十分大きい時、係数の推定値の幾つかを丁度0にさせる効果がある.
- よって、最良部分集合選択のように、lassoは変数選択を実行する.
- lassoはスパースなモデル、つまり変数のある部分集合のみを含むようなモデルを作り出す.
- リッジ回帰のように、lassoでもλの良い値を選択する事は重要である. ここでも交差検証が用いられる.

例: クレジットデータ


標準化された係数値
400
300
200
100
0
−200
20
50
100
200
500
2000
5000
λ


標準化された係数値
400
300
200
100
0
−100
−300
Income
Limit
Rating
Student
0.0
0.2
0.4
0.6
0.8
1.0
$||\hat{\beta}_\lambda^L||_1/||\hat{\beta}||_1$

Lassoの変数選択の性質

なぜlassoは、リッジ回帰と違って、係数の推定値を丁度0にするのか？

Lassoの変数選択の性質

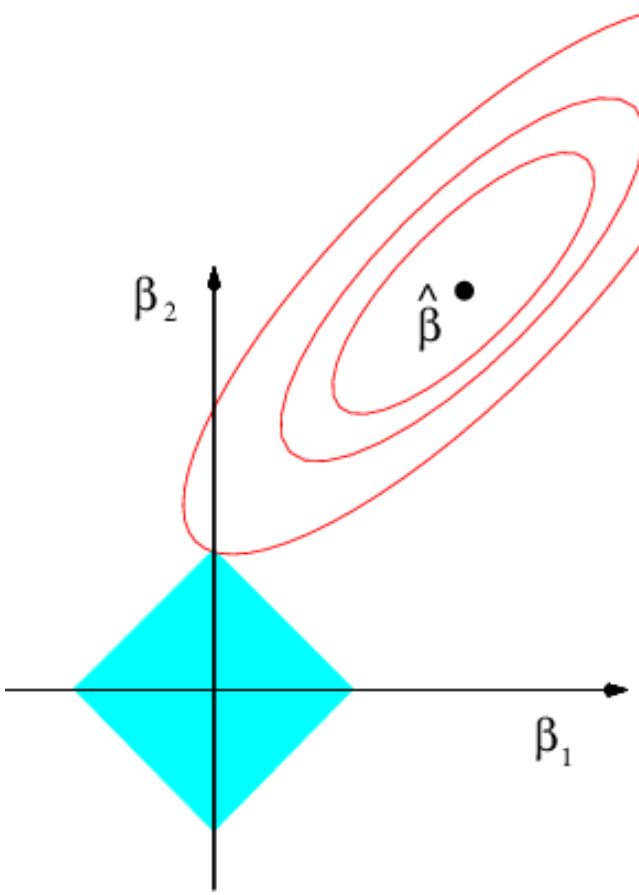
なぜlassoは、リッジ回帰と違って、係数の推定値を丁度0にするのか？
lassoやリッジ回帰による係数推定値はそれぞれ以下の問題を解く事によって得られる事が分かる.

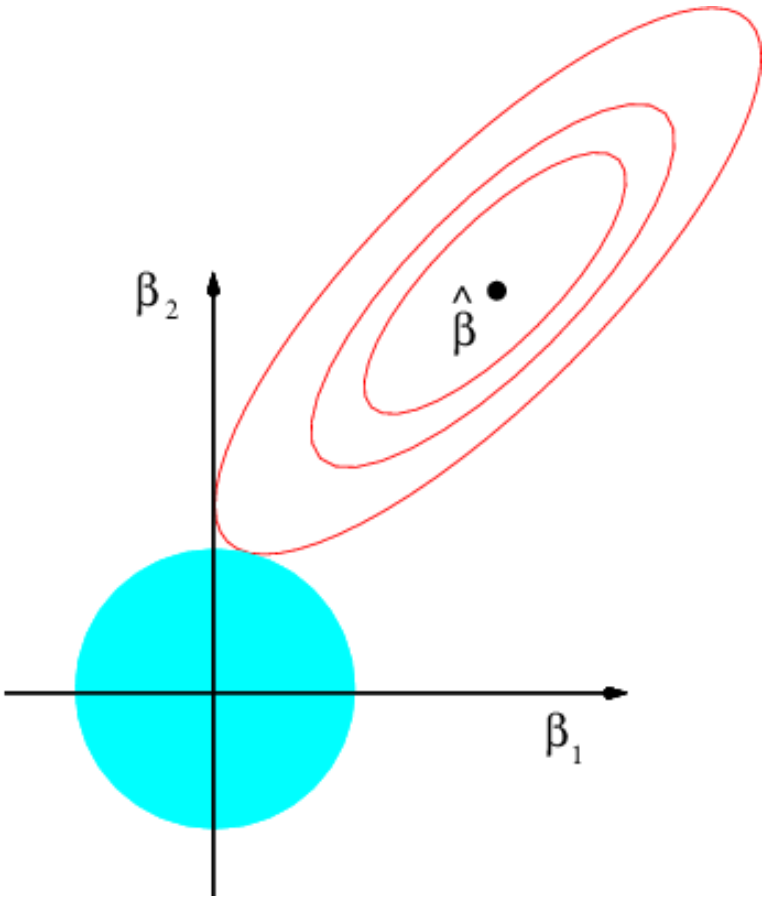
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

と

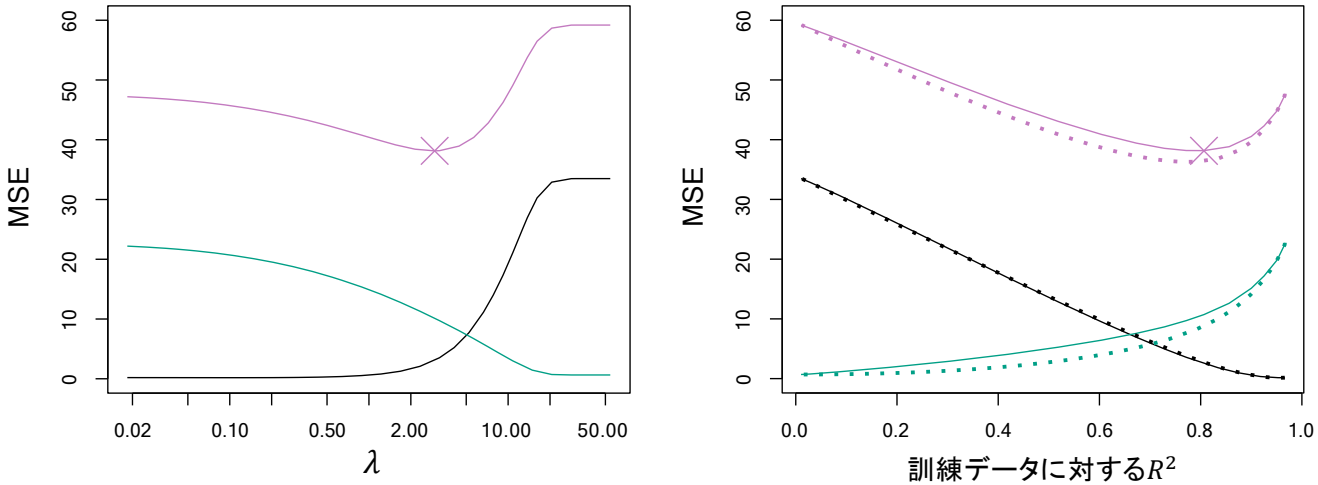
$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq s.$$

Lassoの図





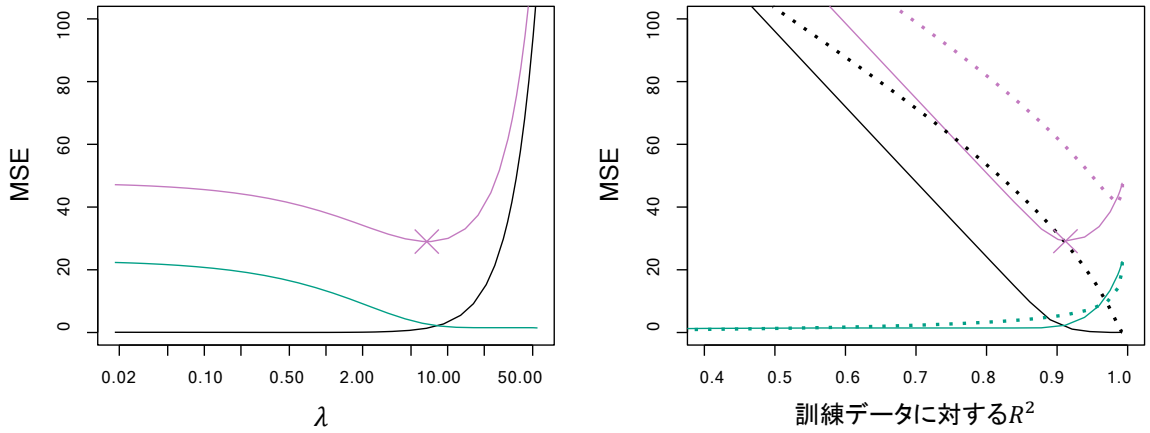
Lassoとリッジ回帰の比較



左: スライド32のシミュレーションデータに対するlassoの、二乗バイアス(黒)、分散(緑)、テストMSE(紫).

右: lasso(実線)とリッジ(点線)の二乗バイアス、分散、テストMSEの比較. 訓練データに対するR^2に対してプロットしている. どちらの図でも×はMSEを最小にするlassoモデルを示している.

Lassoとリッジ回帰の比較: 続き



左: lassoの、二乗バイアス(黒)、分散(緑)、テストMSE(紫).スライド38のシミュレーションデータと似たものだが、応答には2つの予測変数のみが関連している.

右: lasso(実線)とリッジ(点線)の二乗バイアス、分散、テストMSEの比較. 訓練データに対するR^2に対してプロットしている. どちらの図でも×はMSEを最小にするlassoモデルを示している.

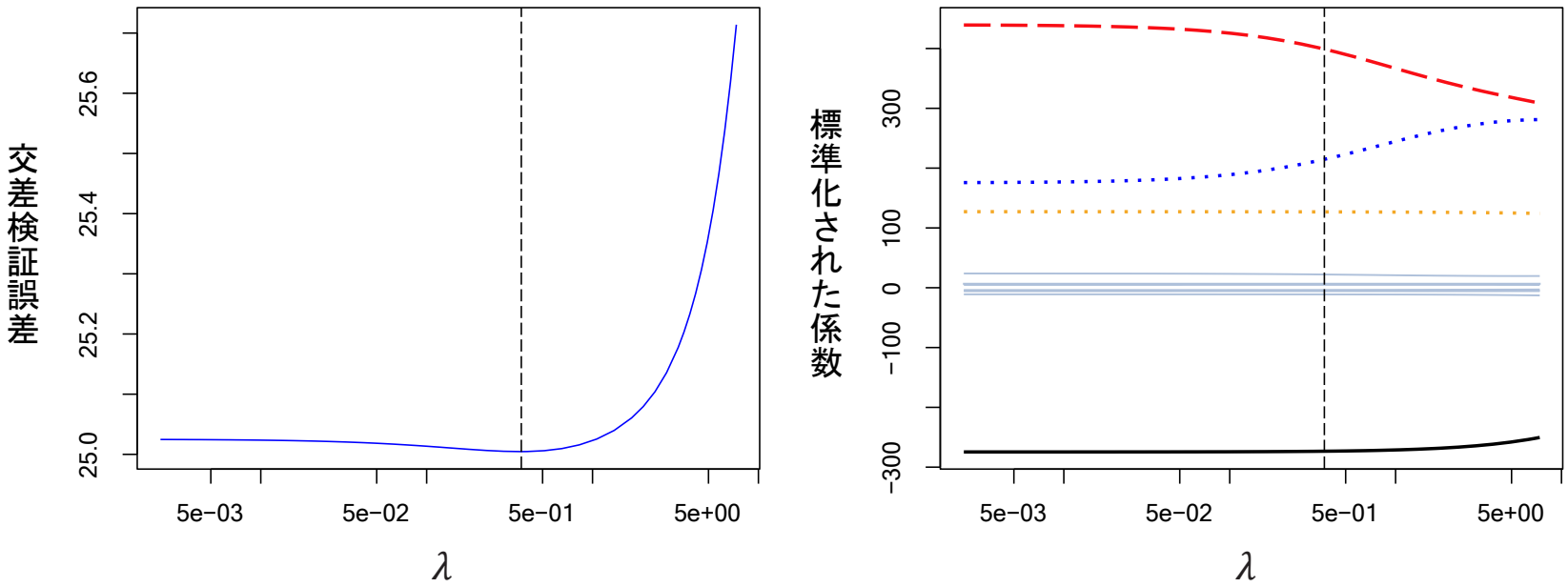
結論

- これらの2つの例はリッジ回帰とlassoのどちらも他を優越する訳ではない事を示している.
- 一般に、応答が比較的少ない予測変数の関数である時、lassoがより良いパフォーマンスを示す事が期待される.
- しかし、現実のデータセットに対しては、応答に関連する予測変数の数は前もっては何も知らない.
- 交差検証のようなテクニックは、ある特定のデータセットに対してより良い方法を決定するために用いられ得る.

リッジ回帰とLassoのチューニングパラメータの選択

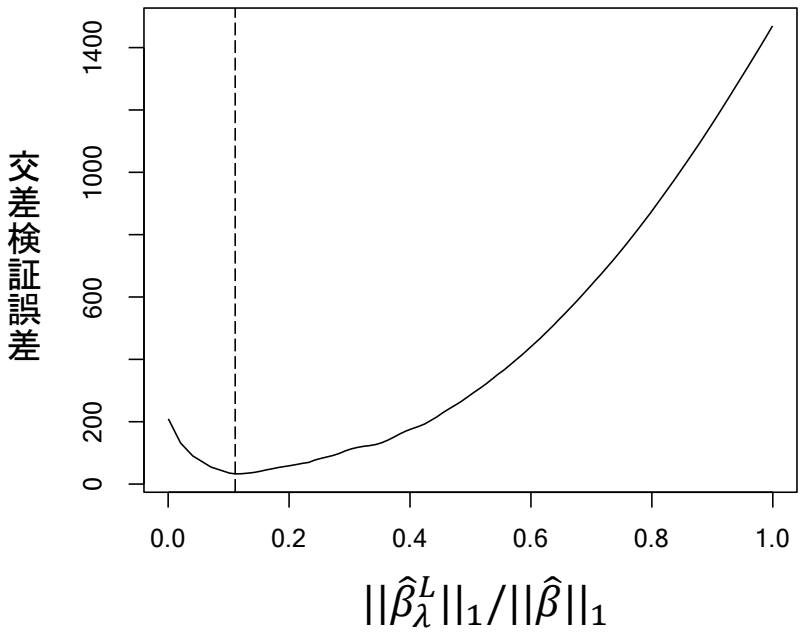
- 部分集合選択と同様に、リッジ回帰とlassoに対しては、どのモデルが最良であるか決定するための方法が必要である.
- つまり、チューニングパラメータλの値、制約の値sを選択する方法が必要である.
- 交差検証は、この問題に取り組む簡単な方法を与える. λの値をグリッドで選び、各値に対する交差検証誤差率を計算する.
- 交差検証誤差を最小にするチューニングパラメータを選ぶ.
- 最後に、使える全ての観測と選ばれたチューニングパラメータの値を用いてモデルを再度あてはめる.

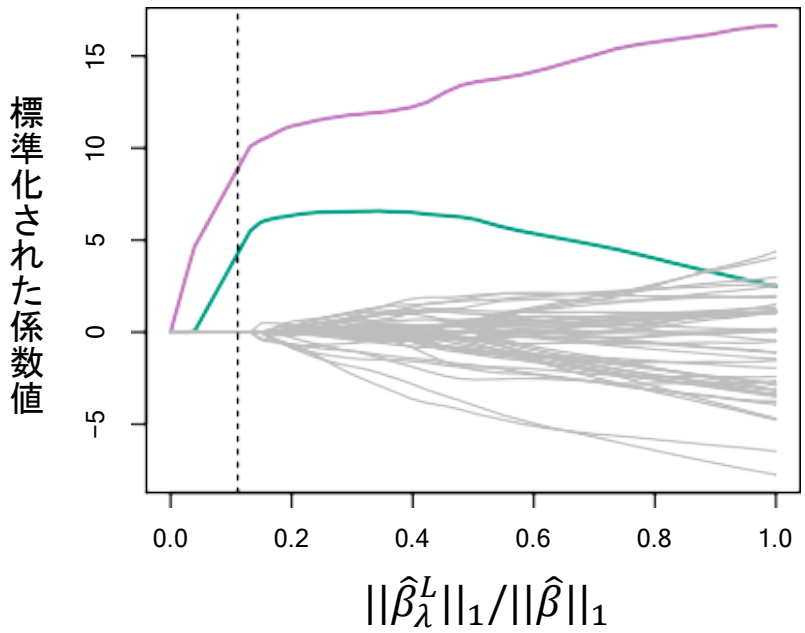
クレジットデータの例



左: クレジットデータセットに対して、λの様々な値でリッジ回帰を適用した際の交差検証誤差.
右: λの関数としての係数の推定値. 縦向きの点線は交差検証で選ばれたλの値を示している

シミュレーションデータの例





左: lassoに対する10分割交差検証MSE、スライド39のスパースなシミュレーションデータに適用した.

右: 対応するlassoによる係数の推定値を示している. 縦向きの点線は交差検証誤差を最小にするlassoの値を示している.

次元削減法

- この章で今まで議論してきた方法は、もともとの予測変数$X_1, X_2, \ldots, X_p$を使った最小二乗や縮小法を用いた、線形回帰モデルのあてはめを含んでいる.
- 予測変数を変換し、変換した変数に対して最小二乗をあてはめる方法について見て行く. これらの方法は次元削減法と呼ばれる.

次元削減法: 詳細

- $Z_1, Z_2, \ldots, Z_M, \mathrm{M} < \mathrm{p}$により、もとの$p$個の予測変数の線形結合を表す. つまり、ある定数$\phi_{m1}, \ldots, \phi_{mp}$に対して

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \,. \tag{1}$$

- 続いて、最小二乗を用いて線形回帰モデルをあてはめる.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \qquad i = 1, \ldots, n. \tag{2}$$

- モデル(2)では、回帰係数は$\theta_0, \theta_1, \ldots, \theta_M$によって与えられている. 定数$\phi_{m1}, \ldots, \phi_{mp}$が上手く選ばれていれば、このような次元削減法は、最小二乗回帰を上回り得る.

- (1)の定義から

$$\sum_{m=1}^{M}\theta_m z_{im}=\sum_{m=1}^{M}\theta_m\sum_{j=1}^{p}\phi_{mj}x_{ij}=\sum_{j=1}^{p}\sum_{m=1}^{M}\theta_m\phi_{mj}x_{ij}=\sum_{j=1}^{p}\beta_j x_{ij}\,.$$

ただし

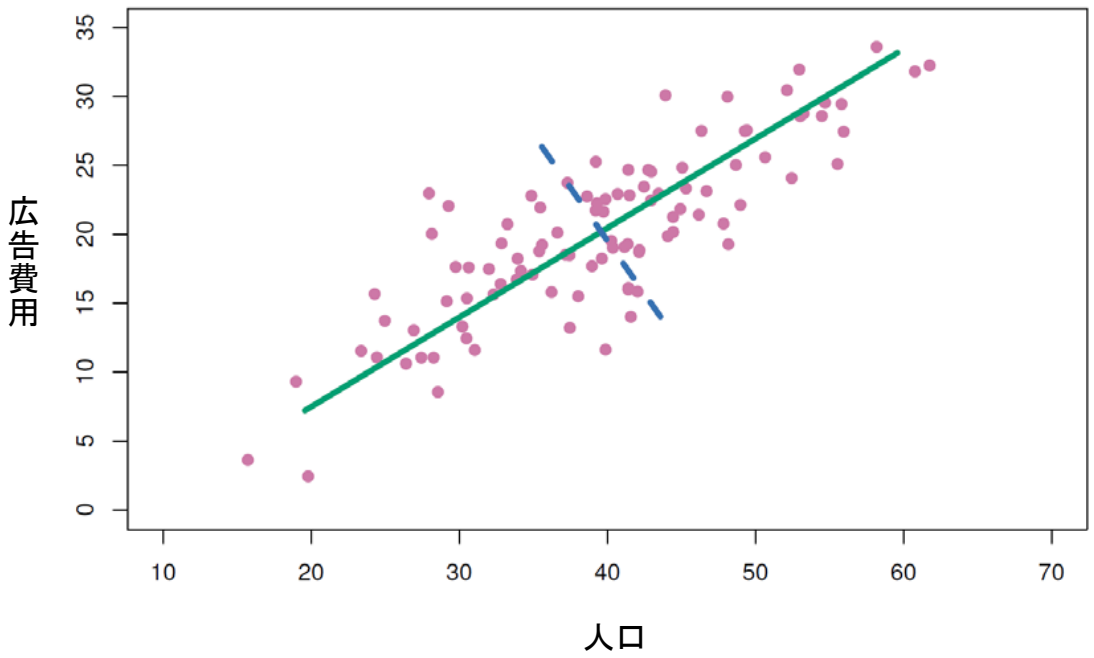
$$\beta_j=\sum_{m=1}^{M}\theta_m\phi_{mj}\,. \qquad (3)$$

- よってモデル(2)は最小二乗回帰の特殊ケースと考える事が出来る.
- (3)の形から、次元削減は係数β_jの推定値を制約する作用がある.
- バイアス-分散のトレードオフを上手く対処しうる.

主成分回帰

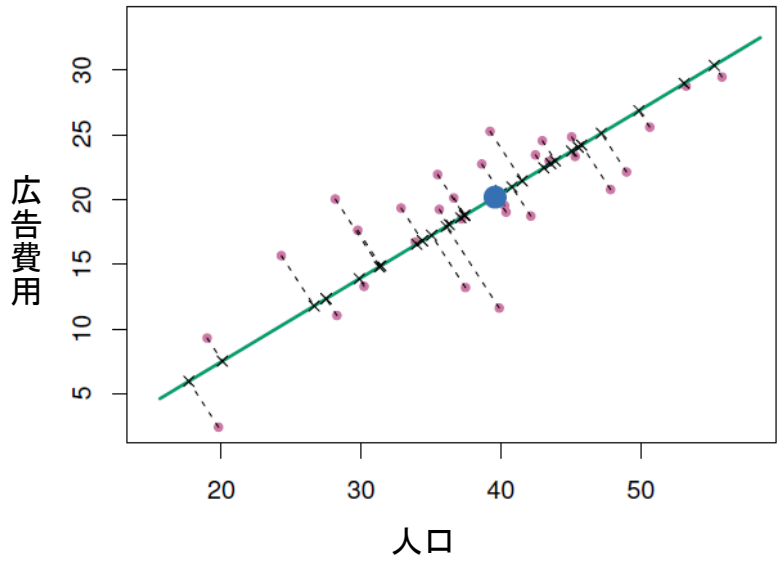
- ここでは主成分分析(PCA)(テキストの12章で議論される)を予測変数の線形結合を定めるために用いる. そして回帰を行う.
- 第1主成分は、変数の(標準化された)線形結合の内の分散が最大となるもの.
- 第2主成分は、第1主成分と無相関なものの内、変数の(標準化された)線形結合の内の分散が最大となるもの.
- 続く.
- 相関をもった変数が多くあるとき、共通の変動を捉える主成分の小さな集合で置き換える.

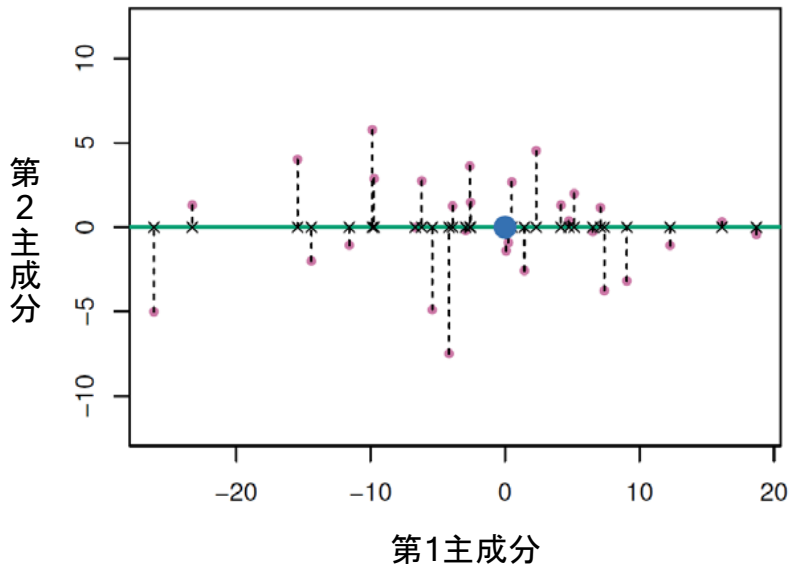
PCAの図



100都市の人口の大きさ(pop)と広告費用(ad)が紫の点で示されている. 緑の太線は第1主成分、青の破線は第2主成分を示している.

PCAの図: 続き



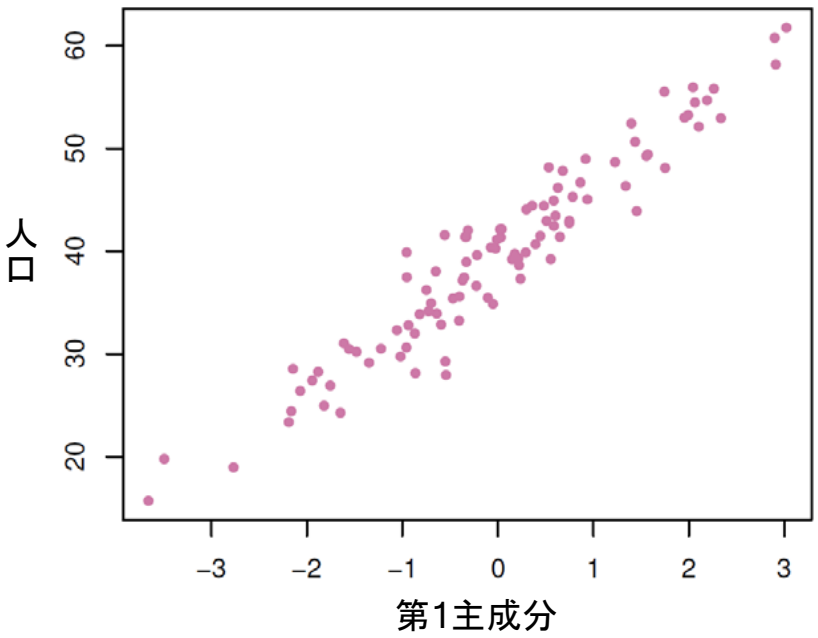


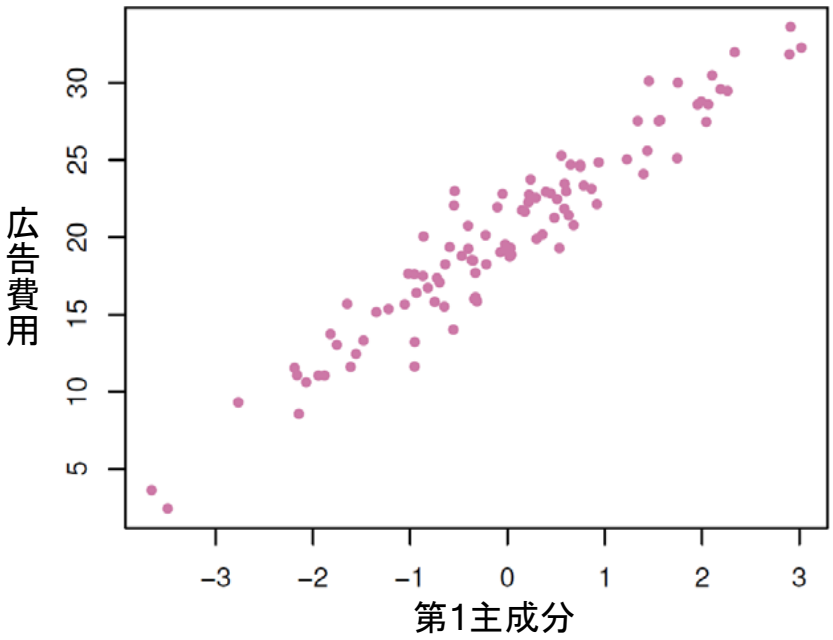
広告データの抜き出し.

左: 第1主成分は、各点からの垂線の2乗距離の和が最小となるように選ばれ、緑で書かれている. これらの距離は黒の破線で表されている.

右: 右図は第1主成分がx軸となるように回転したもの.

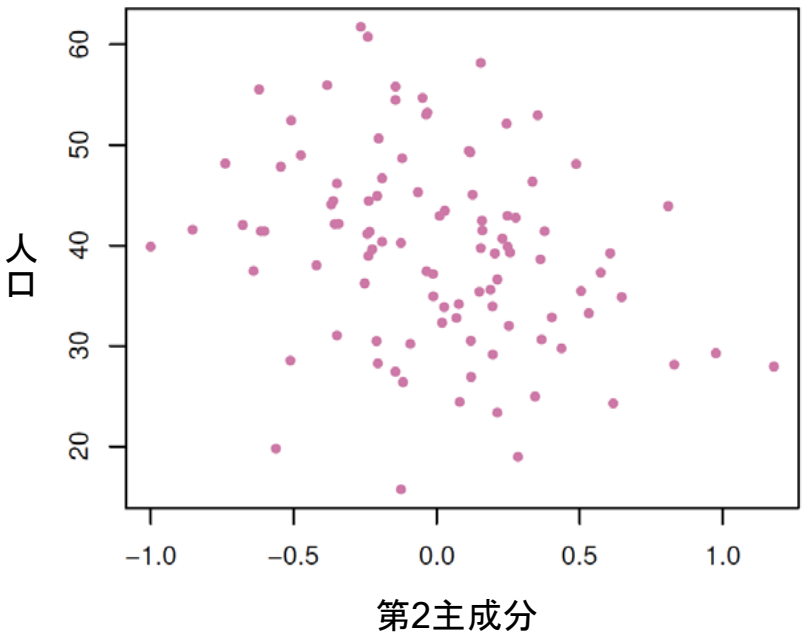
PCAの図: 続き

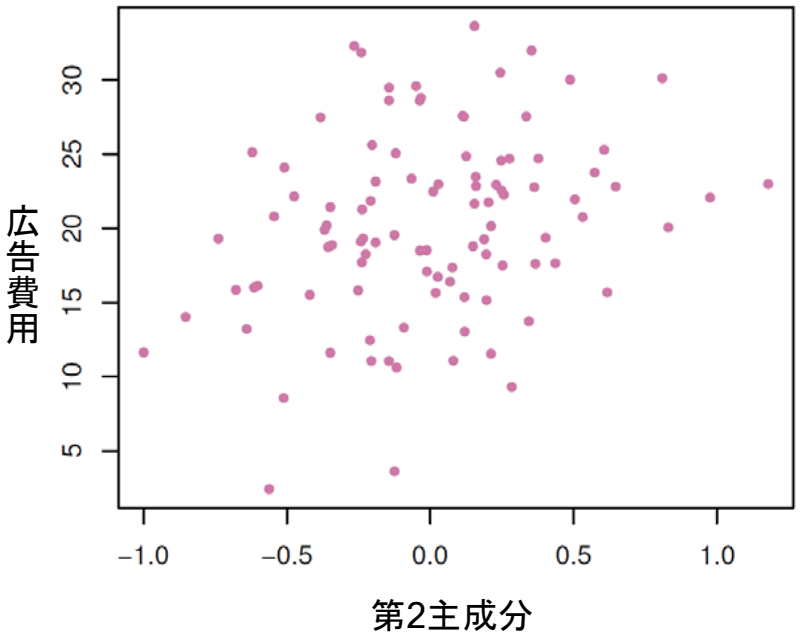




第1主成分スコアz_{i1}とpop, adのプロット. 相関が強くなっている.

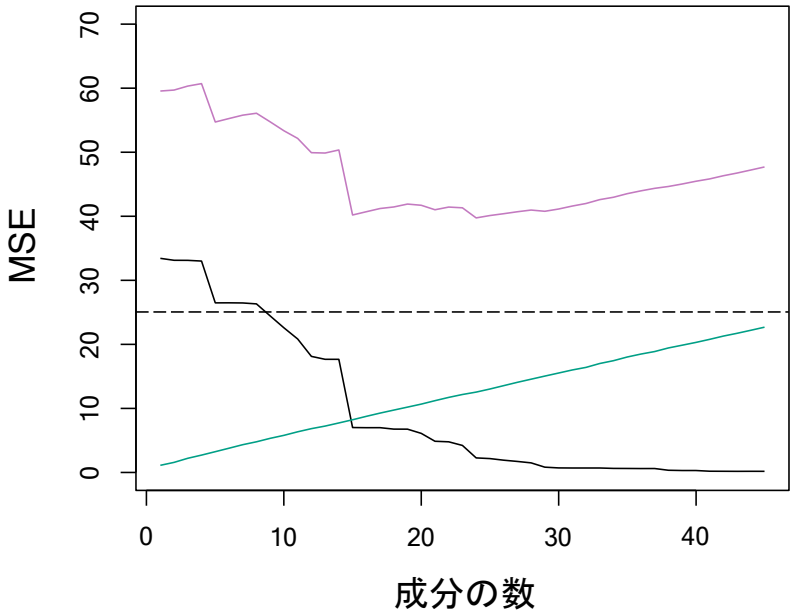
PCAの図: 続き

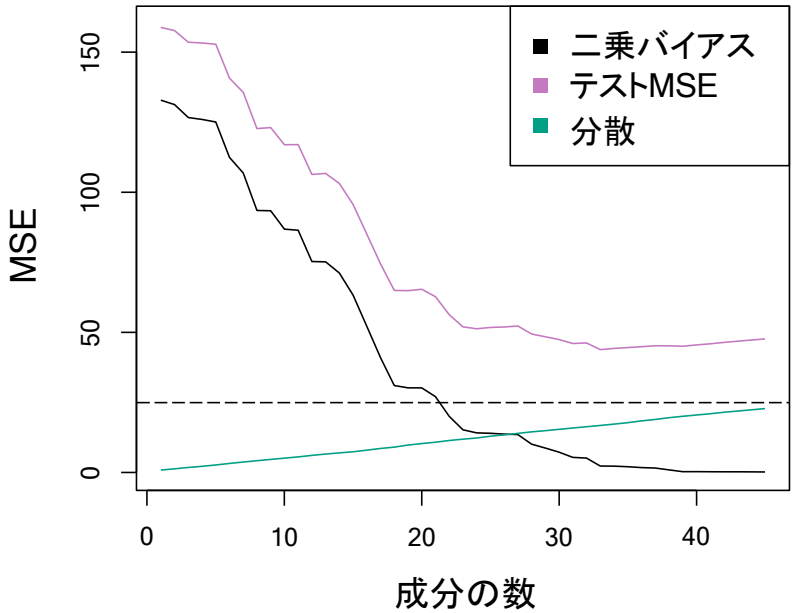




第2主成分スコアz_{i2}とpop, adのプロット. 相関が弱くなっている.

主成分回帰の適用



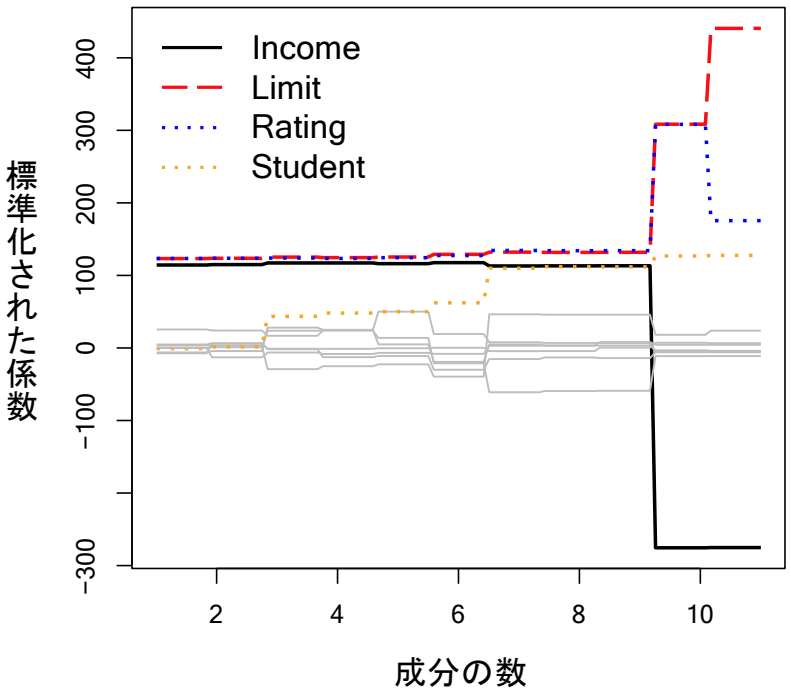


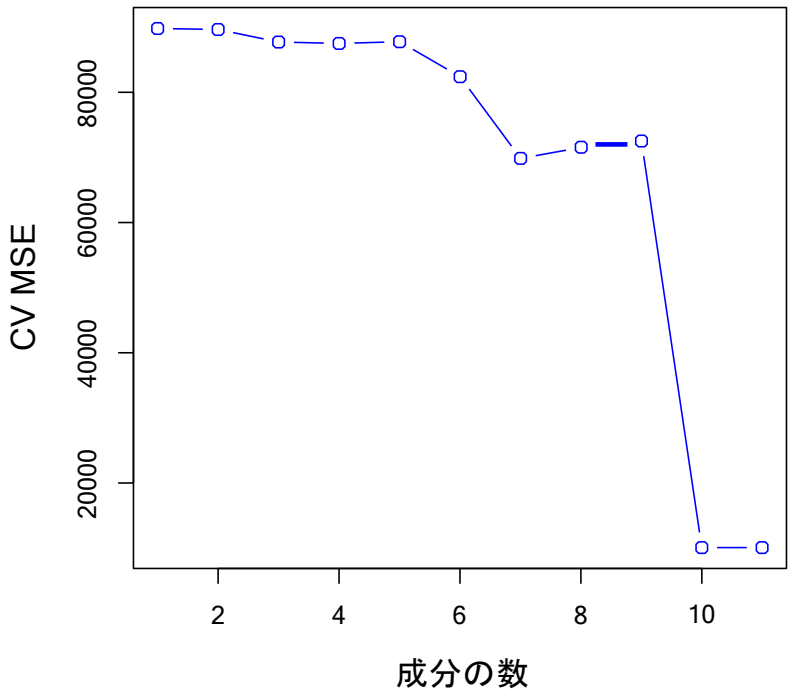
主成分回帰(PCR)を2つのシミュレーションデータに適用した. 黒線、緑線、紫線は二乗バイアス、分散、テストMSEに対応している.

左: スライド32のシミュレーションデータ.

右: スライド39のシミュレーションデータ.

方向の数Mの選択





左: クレジットデータセットに対し、異なる値のMを用いたときのPCRによる標準化された係数の推定値.

右: Mの関数として、PCRを用いて得られた10-分割交差検証MSE

部分最小二乗

- PCRは予測変数$X_1, \ldots, X_p$を最も良く表現する線形結合, 方向を特定する.
- 応答Yは主成分の方向を定めるためには用いられないので、これらの方向は教師なしの方法で特定される.
- つまり、応答は主成分の特定に役立てられていない.
- 結果として、PCRは潜在的に深刻な欠陥を持っている. 予測変数をよく説明するような方向が、応答の予測のために用いられる最良の方向でもあるという保証はない.

部分最小二乗: 続き

- PCRのように、PLSは次元削減法の1つである. まずもとの特徴の線形結合による特徴の集合$Z_1, \ldots, Z_M$を特定する. さらに、これらのM個の特徴を用いて最小二乗による線形モデルのあてはめを行う.
- しかしPCRと違って、PLSはこれらの特徴を教師ありの方法で定める. つまり、応答Yを役立てて、新たな特徴がもとの特徴を良く近似するだけでなく、応答に関連するようにも定める.
- 雑に言うと、PLSは応答と予測変数の両方を説明するような方向を見つけようとする.

部分最小二乗の詳細

- p個の予測変数を標準化した後、PLSは、(1)でのϕ_{1j}を、YのX_jへの線形単回帰による係数に等しくなるように定め、1つめの方向Z_1を計算する.
- この係数がYとX_jの相関に比例することが示せる.
- $Z_1 = \sum_{j=1}^{p} \phi_{1j} X_j$を計算することで、PLSは応答に最も強く関連するような変数に大きな重みが置かれる.
- 続く方向は、残差に対し、上記を繰り返す事で得られる.

まとめ

- モデル選択法はデータ分析、中でも多くの予測変数を含むビッグデータでの本質的なツールである.
- 例えばlassoのような、スパース性を与えるような方法に関する研究は中でも盛り上がっている.
- 後に、スパース性についてより詳細に触れる. elastic net のような関連する方法についても説明する.

elastic net (訳者注)

- elastic netの説明はISLにはない.
- elastic netによる係数推定値$\hat{\beta}$は次を最小化する

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left(\alpha\beta_j^2+(1-\alpha)\left|\beta_j\right|\right)$$

$$=RSS+\lambda\sum_{j=1}^{p}\left(\alpha\beta_j^2+(1-\alpha)\left|\beta_j\right|\right)$$

ただし$\lambda\geq 0, 1\geq\alpha\geq 0$はチューニングパラメータで、それぞれの場合で決定される.

- LASSO, 縮小回帰の一般化と言える. 詳しくはHastie-Tibshirani-FriedmanのESLを参照.

第7章: 線形性からの逸脱
-Moving Beyond Linearity-

- 真の世界は線形ではない or ほとんどの場合は線形でない
- しばしば線形性の仮定は十分に役に立つ
- ロジスティック回帰-Logistic regression-
- 多項式-polynomials-
- 階段関数-step functions-
- スプライン-splines-
- 局所回帰-local regression-
- 一般加法モデル-generalized additive models-

第7章: 線形性からの逸脱
-Moving Beyond Linearity-

真の世界は線形ではない!

第7章: 線形性からの逸脱
-Moving Beyond Linearity-

真の世界は線形ではない!
あるいはほとんどの場合は線形でない!

第7章: 線形性からの逸脱
-Moving Beyond Linearity-

真の世界は線形ではない!
あるいはほとんどの場合は線形でない!

しかしながらしばしば線形性の仮定は十分に役に立つ.

第7章: 線形性からの逸脱
-Moving Beyond Linearity-

真の世界は線形ではない!
あるいはほとんどの場合は線形でない!

しかしながらしばしば線形性の仮定は十分に役に立つ.

線形ではないと...

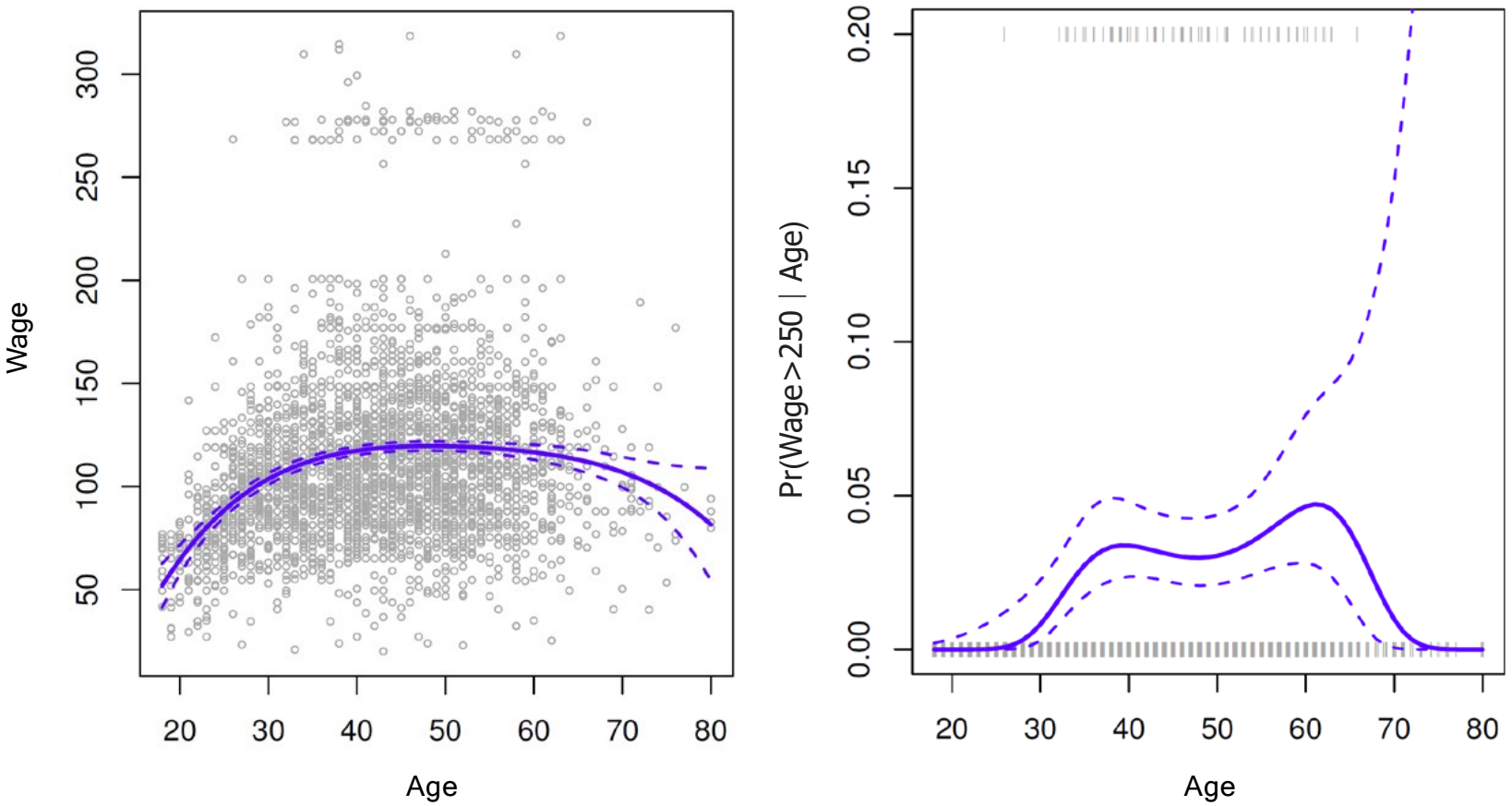
- 多項式polynomials,
- 階段関数step functions,
- スプラインsplines,
- 局所回帰local regression,
- 一般加法モデルgeneralized additive models

などはかなり柔軟性があり, それほど扱いが困難ではなく, 線形モデルの解釈可能性を失わない.

多項式回帰(Polynomial Regression)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$$

4次曲線(Degree−4 Polynomial)



詳細

- 新たに変数を作成 $X_1 = X, X_2 = X^2$,etc.として重回帰モデルとする.

詳細

- 新たに変数を作成 $X_1 = X, X_2 = X^2$,etc.として重回帰モデルとする.

- 係数自身にはそれほど関心がなく, 任意の点x_0におけるフィット値により関心がある:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4.$$

詳細

- 新たに変数を作成 $X_1 = X, X_2 = X^2$,etc.として重回帰モデルとする.

- 係数自身にはそれほど関心がなく, 任意の点x_0におけるフィット値により関心がある:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4.$$

- ここで $\hat{f}(x_0)$は係数$\hat{\beta}_\ell$の線形関数なので任意の点 x_0における各点分散(*pointwise-variances*) $\mathrm{Var}[\hat{f}(x_0)]$が容易に求まる. 図に中でフィットした関数と各点の標準誤差を計算, $\hat{f}(x_0) \pm 2 \cdot \mathrm{se}[\hat{f}(x_0)]$ を示しておく.

詳細

- 新たに変数を作成 $X_1 = X, X_2 = X^2$,etc.として重回帰モデルとする.

- 係数自身にはそれほど関心がなく, 任意の点x_0におけるフィット値により関心がある:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4.$$

- ここで $\hat{f}(x_0)$は係数$\hat{\beta}_\ell$の線形関数なので任意の点 x_0における各点分散(*pointwise-variances*) $\mathrm{Var}[\hat{f}(x_0)]$が容易に求まる. 図に中でフィットした関数と各点の標準誤差を計算, $\hat{f}(x_0) \pm 2 \cdot \mathrm{se}[\hat{f}(x_0)]$ を示しておく.

- 次数dをある適切に固定しておくか, 交差検証法(クロスバリデーション,cross-validation)によりdを選択する.

続き

- ロジスティック回帰(Logistic regression)が自然と導かれる. 例えば, 図は統計モデル :

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}$$

- 信頼区間を得るには, ロジットスケールで上界, 下界を計算, 確率値に反転すればよい.

続き

- ロジスティック回帰(Logistic regression)が自然と導かれる. 例えば, 図は統計モデル :

$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}$$

- 信頼区間を得るには, ロジットスケールで上界, 下界を計算, 確率値に反転すればよい.
- 幾つかの変数について別々に計算できるだろうか—変数を行列としてまとめ, あとで各要素を利用する(separate out the pieces)(後でGAMについて述べる).

続き

- ロジスティック回帰(Logistic regression)が自然と導かれる. 例えば, 図は統計モデル :

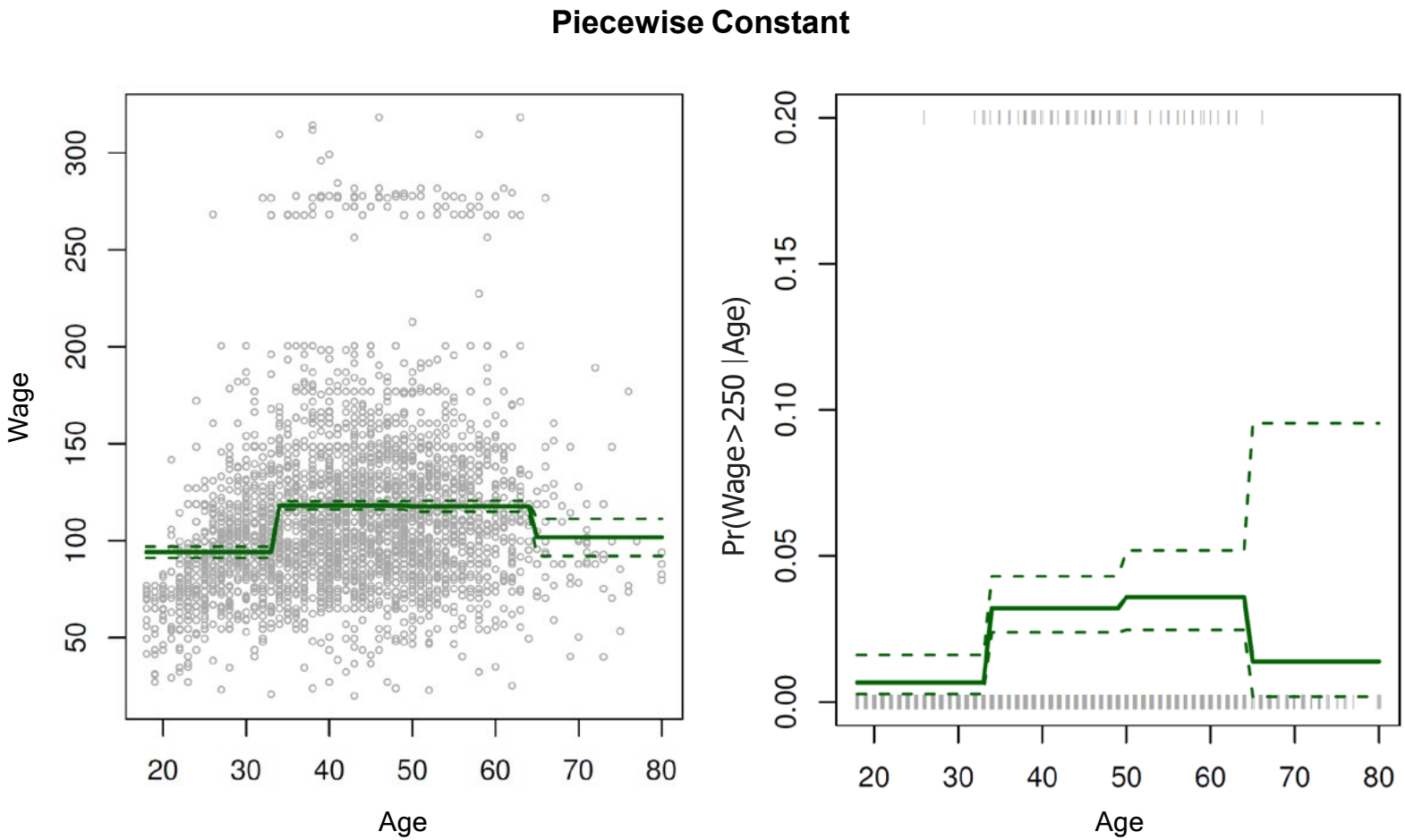
$$\Pr(y_i > 250|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d)}$$

- 信頼区間を得るには, ロジットスケールで上界, 下界を計算, 確率値に反転すればよい.
- 幾つかの変数について別々に計算できるだろうか—変数を行列としてまとめ, あとで各要素を利用する(separate out the pieces)(後でGAMについて述べる).
- 注意Caveat: 多項式は悪名高い裾の挙動がある — 外挿には悪い結果をもたらす.
- コマンド y ~ poly(x, degree = 3) によりフィットできる.

階段関数(Step Functions)

変数を変換するもう一つの方法は— 各領域で変数をカットする.

$C_1(X) = I(X < 35), C_2(X) = I(35 \leq X < 50), \ldots, C_3(X) = I(X \geq 65)$



階段関数, 続き

- 容易に適用できる. 各グループを表現するダミー変数列を作る.

階段関数, 続き

- 容易に適用できる. 各グループを表現するダミー変数列を作る.
- 解釈が容易な交互作用を表現する有用で方法. 例えば年と年齢Year, Ageの交互作用:

$$I(\text{Year} < 2005) \cdot \text{Age}, \qquad I(\text{Year} \geq 2005) \cdot \text{Age}$$

とすると各年齢層に異なる線形関数が与えられる.

階段関数, 続き

- 容易に適用できる. 各グループを表現するダミー変数列を作る.
- 解釈が容易な交互作用を表現する有用で方法. 例えば年と年齢Year, Ageの交互作用:

$$I(\text{Year} < 2005) \cdot \text{Age}, \qquad I(\text{Year} \geq 2005) \cdot \text{Age}$$

とすると各年齢層に異なる線形関数が与えられる.

- Rプログラムでは:
I(year < 2005) あるいは cut(age, c(18, 25, 40, 65, 90))
を利用する.

階段関数, 続き

- 容易に適用できる. 各グループを表現するダミー変数列を作る.
- 解釈が容易な交互作用を表現する有用で方法. 例えば年と年齢Year, Ageの交互作用:

$$I(\text{Year} < 2005) \cdot \text{Age}, \qquad I(\text{Year} \geq 2005) \cdot \text{Age}$$

  とすると各年齢層に異なる線形関数が与えられる.
- Rプログラムでは:
  I(year < 2005) あるいは cut(age, c(18, 25, 40, 65, 90))
  を利用する.
- 境界点, 結節点(cutpoints, *knots*)は問題がある. 非線形性を作る滑らかな方法としてはスプライン(*splines*)などがある.

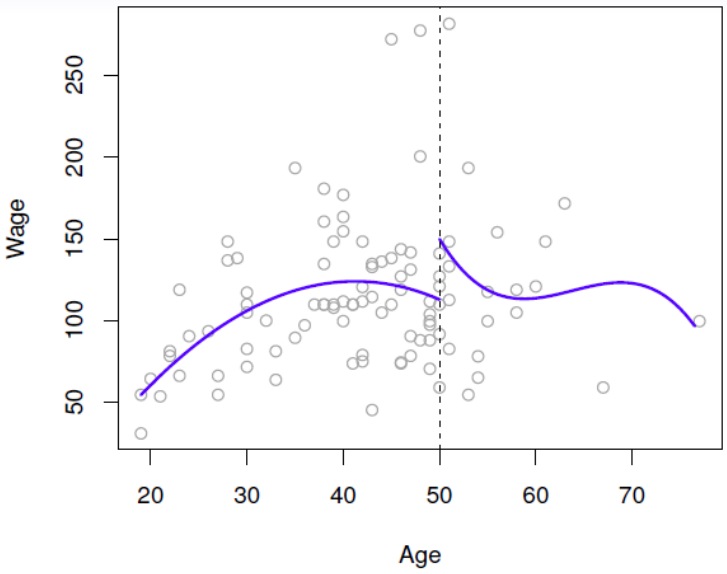
区分的多項式
(Piecewise Polynomials)

- すべての領域における変数xの多項式の代わりに結節点で定義された各領域で異なる多項式を用いることができる.
  例えば (図を参照)

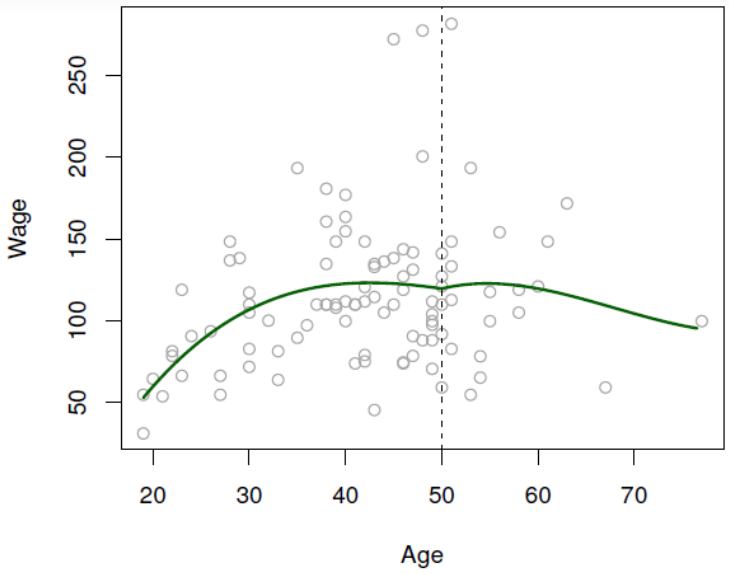
$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

- ここで多項式に制約を加えた方が良い, 例えば
  連続性etc.
- *スプライン(Splines)*はもっとも連続性があると言える.

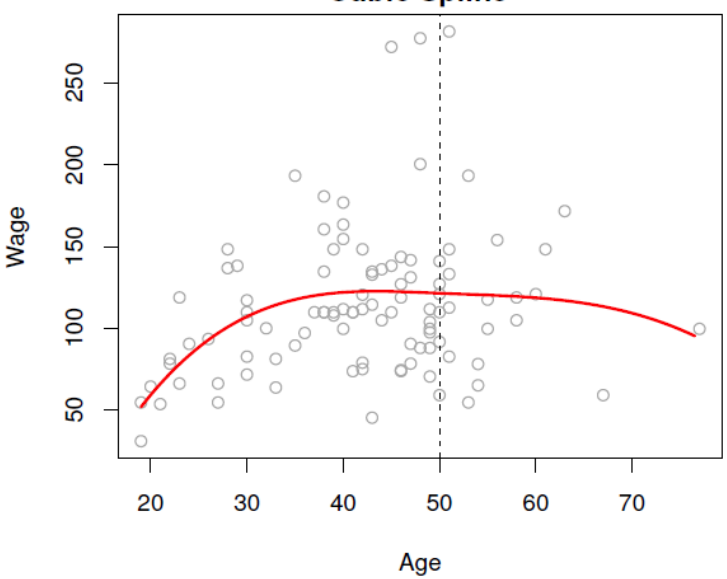
区分3次
Piecewise Cubic



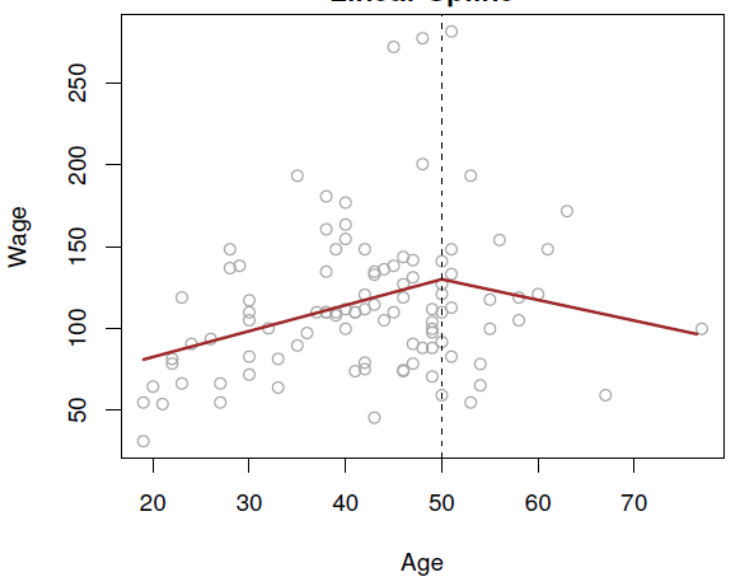
連続的区分3次
Continuous Piecewise Cubic



3次スプライン
Cubic Spline



線形スプライン
Linear Spline



線形スプライン-Linear Splines-

結節点$\xi_k, k = 1, \ldots, K$の線形スプライン(linear spline)は
各結節点で連続な区分線形多項式である.

この統計モデルは次のように表現できる:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

ここでb_kは基本関数(*basis functions*)である.

線形スプライン-Linear Splines-

結節点$\xi_k, k = 1, \ldots, K$の線形スプライン(linear spline)は
各結節点で連続な区分線形多項式である.

この統計モデルは次のように表現できる:

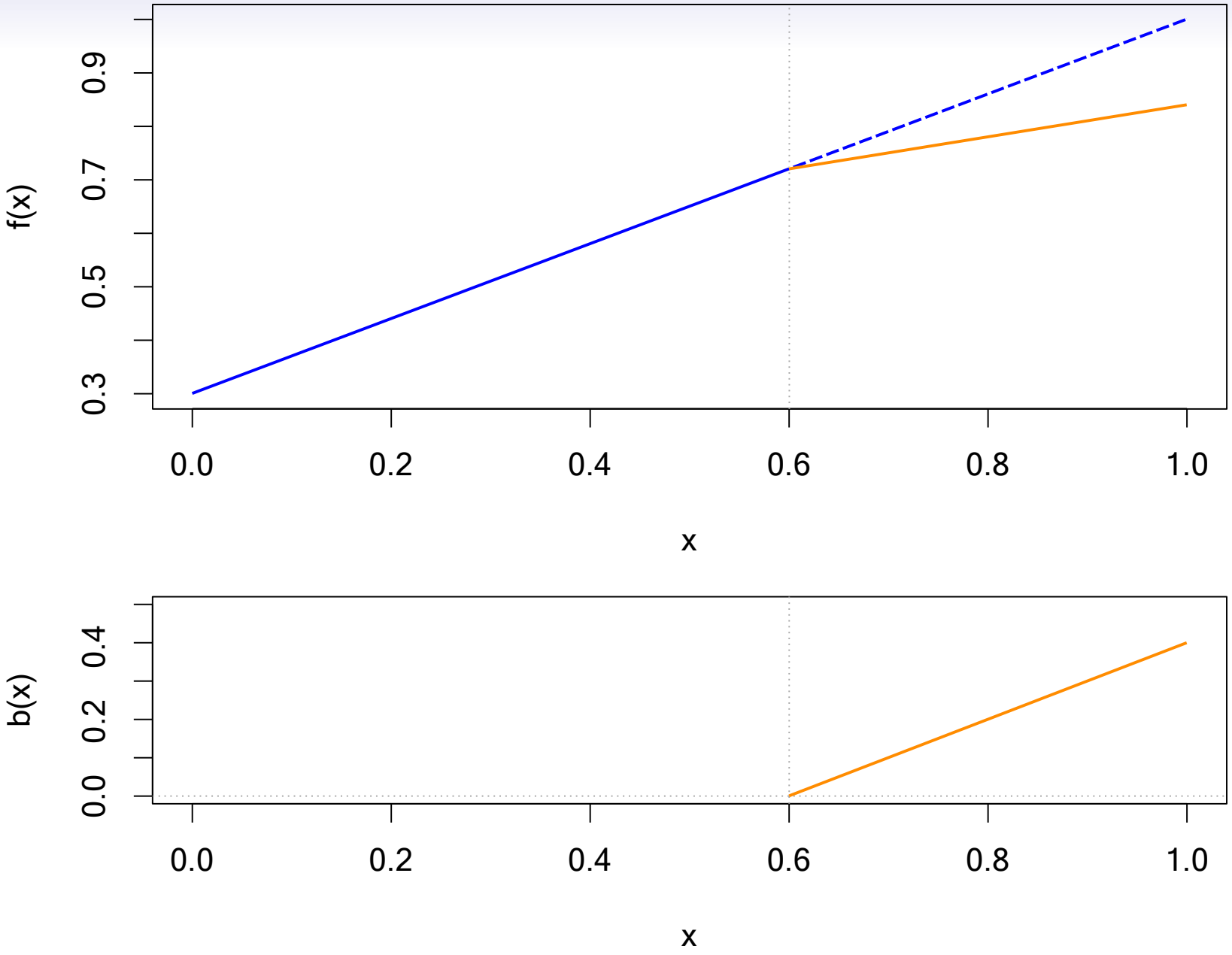
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

ここでb_kは基本関数(*basis functions*)である.

$$b_1(x_i) = x_i$$
$$b_{k+1}(x_i) = (x_i - \xi_k)_+, k = 1, \ldots, K$$

記号$()_+$は正値部分を意味する; つまり

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{他} \end{cases}$$


f(x)
0.3
0.5
0.7
0.9
0.0
0.2
0.4
0.6
0.8
1.0
x
b(x)
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
x

3次スプライン-Cubic Splines-

結節点 $\xi_k, k = 1, \ldots, K$ の3次スプラインは各結節点で微分係数が2次まで連続となる区分3次多項式である.

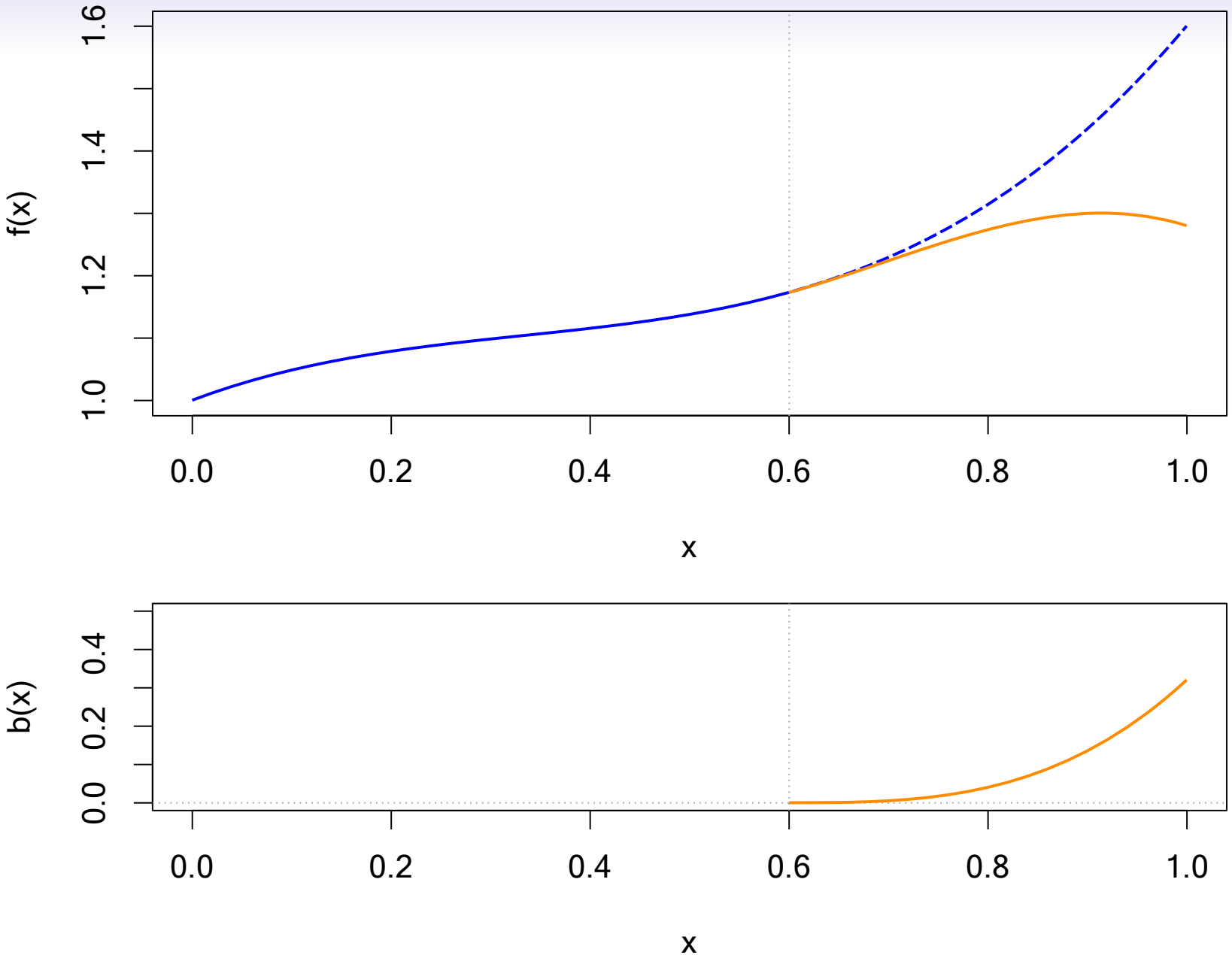
このモデルは切断ベキ関数(truncated power basis functions)により表現できる:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i,$$

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, k = 1, \ldots, K \end{aligned}$$

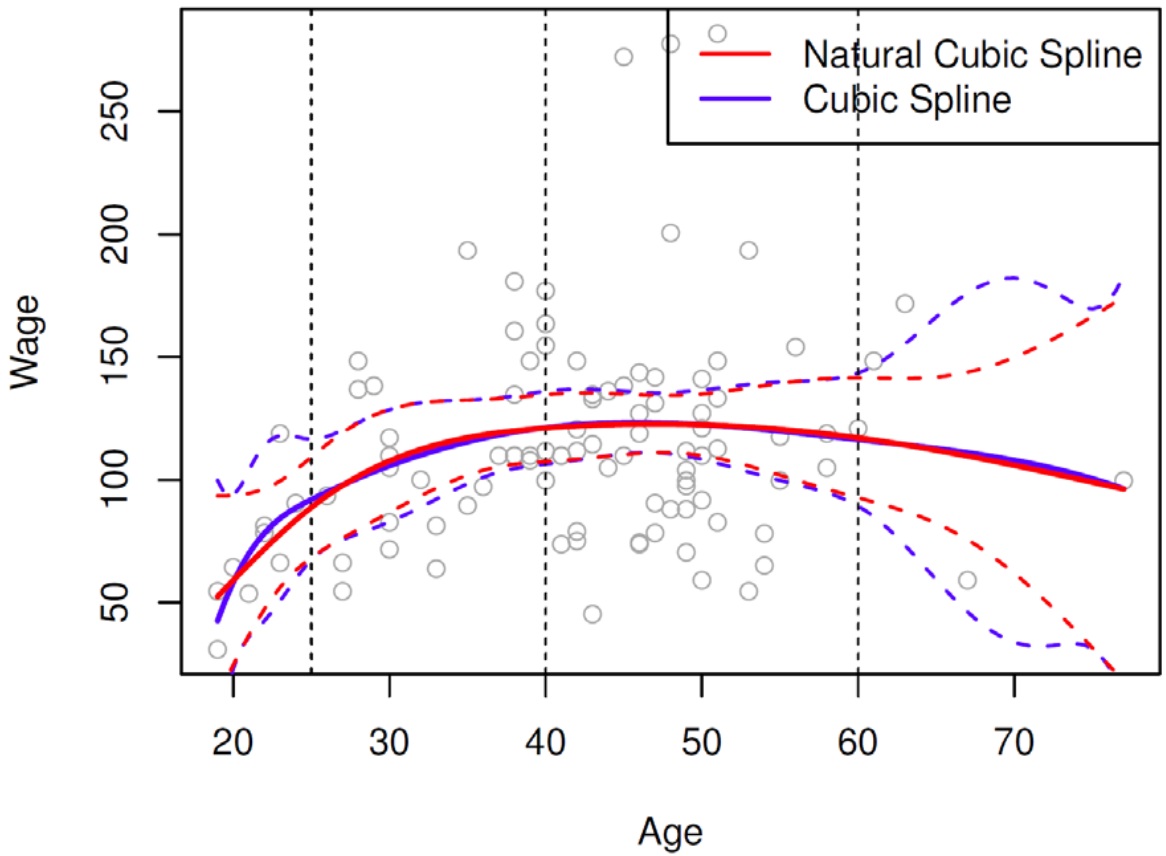
ただし

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i < \xi_k \\ 0 & \text{他} \end{cases}$$

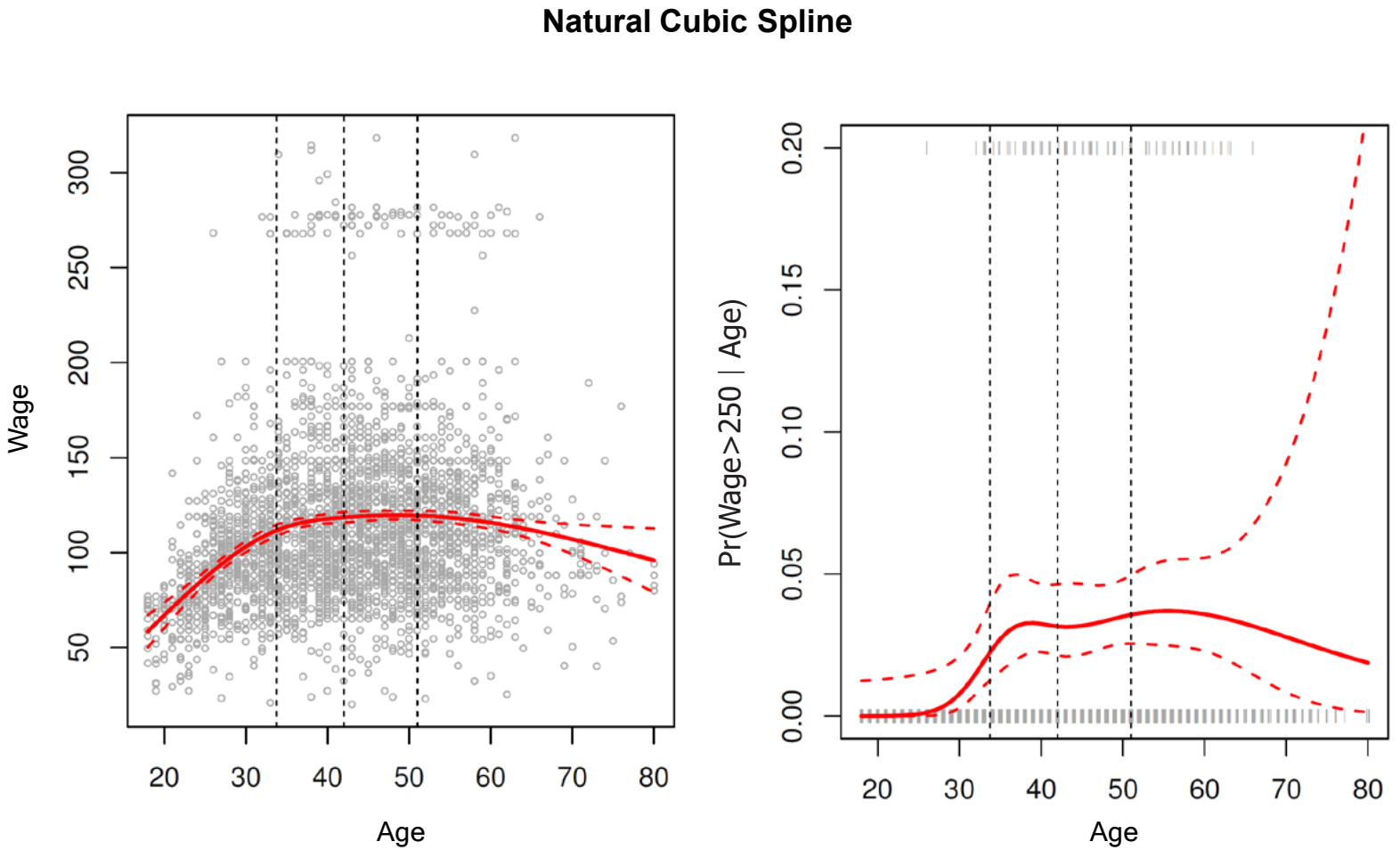

f(x)
1.0
1.2
1.4
1.6
0.0
0.2
0.4
0.6
0.8
1.0
x
b(x)
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
x

自然3次スプライン-Natural Cubic Splines-

自然3次スプライン(natural cubic spline)では有界な結節点の先は線形に外挿する. このことから4 = 2 × 2 個の制約条件が加わり, 通常の3次スプラインより同じ自由度でもより多くの結節点を置くことができる.

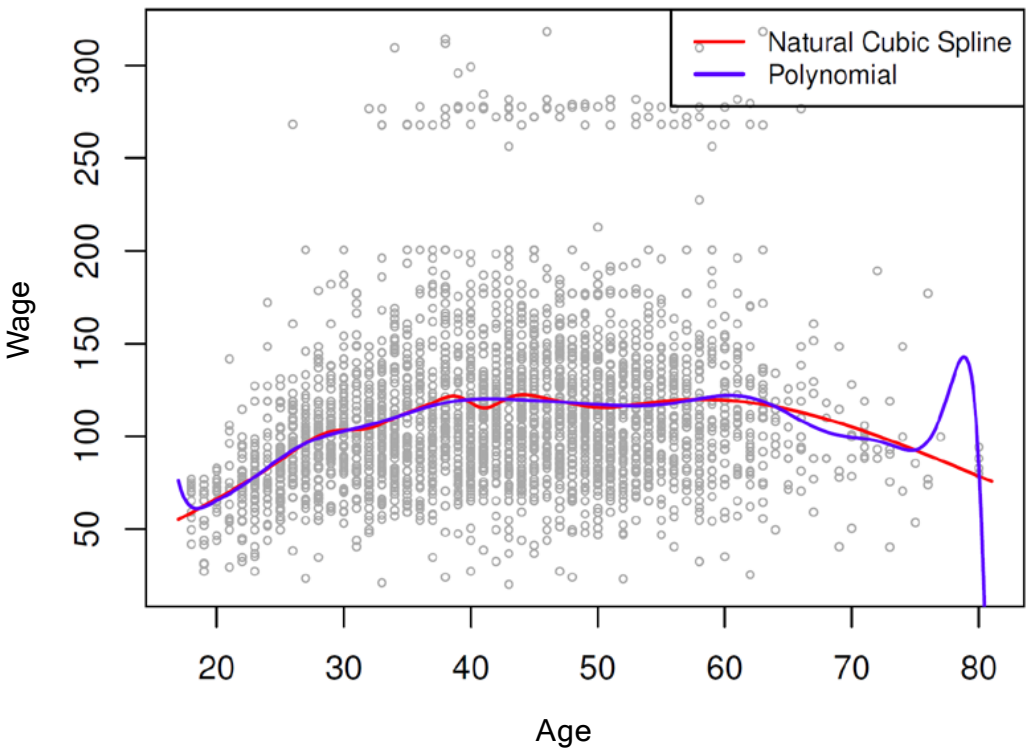


Rを利用スプラインをフィットするにはパッケージsplinesにおいて
任意の次数についてはbs(x, ...):
自然3次スプライン(natural cubic) ns(x, ...) を利用する.



結節点の配置(Knot placement)

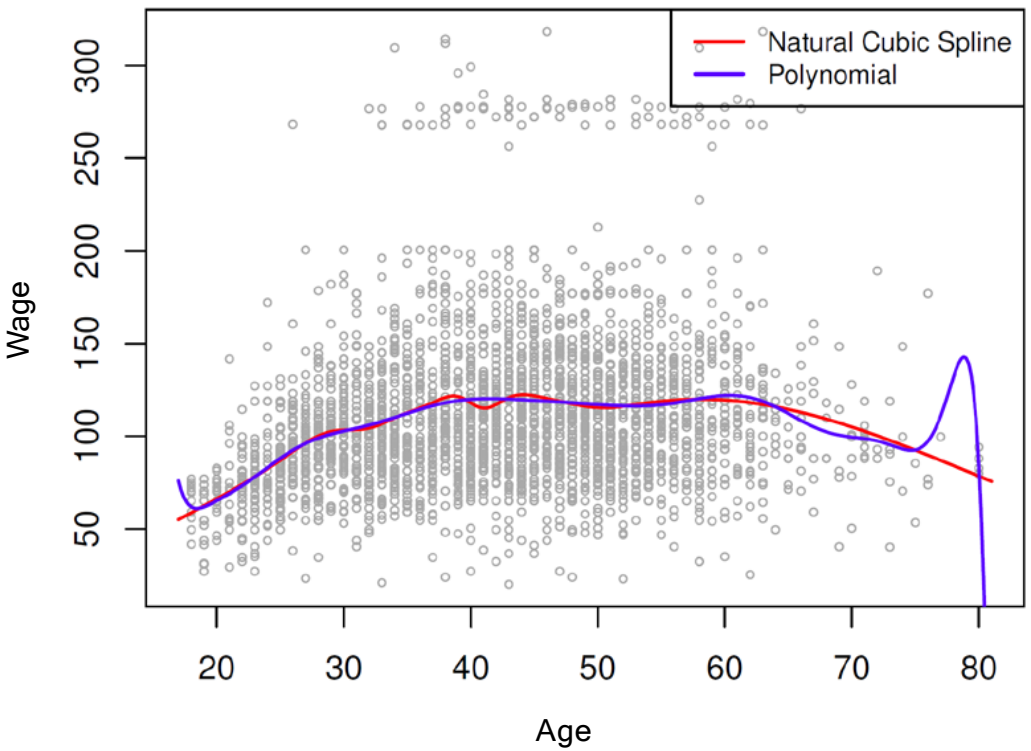
- 結節点の数Kを定め, 次に観測データXの適当な分位点により配置する.
- 結節点Kの3次スプラインは$K+4$個の母数, 自由度を持つ.
- 結節点K個の自然スプライン(natural spline)の自由度はKとなる.



次数14の多項式と
自然スプラインの比較
自由度は15df.

結節点の配置(Knot placement)

- 結節点の数Kを定め, 次に観測データXの適当な分位点により配置する.
- 結節点Kの3次スプラインは$K+4$個の母数, 自由度を持つ.
- 結節点K個の自然スプライン(natural spline)の自由度はKとなる.



次数14の多項式と
自然スプラインの比較
自由度は15df.

```
ns(age,df=14)
poly(age, deg=14)
```

平滑化スプライン-Smoothing Splines-

ここでの議論は少し数理的

滑らかな関数$g(x)$をあるデータにフィットする

次の規準を考える:

$$\underset{g\in\mathcal{S}}{\text{minimize}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

平滑化スプライン-Smoothing Splines-

ここでの議論は少し数理的

滑らかな関数$g(x)$をあるデータにフィットする

次の規準を考える:

$$\underset{g\in\mathcal{S}}{\text{minimize}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda\int g''(t)^2 dt$$

- 第1項はRSSであり,各x_iにおけるデータになるべく$g(x)$ をフィットさせようとしている.

平滑化スプライン-Smoothing Splines-

ここでの議論は少し数理的

滑らかな関数$g(x)$をあるデータにフィットする

次の規準を考える:

$$\underset{g\in\mathcal{S}}{\text{minimize}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- 第1項はRSSであり,各x_iにおけるデータになるべく$g(x)$ をフィットさせようとしている.
- 第2項は粗さのペナルティ(*roughness penalty*)であり,$g(x)$の変動の程度を示す. チューニング母数(*tuning parameter*) $\lambda \geq 0$により制御する.

平滑化スプライン-Smoothing Splines-

ここでの議論は少し数理的

滑らかな関数$g(x)$をあるデータにフィットする

次の規準を考える:

$$\underset{g\in\mathcal{S}}{\text{minimize}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- 第1項はRSSであり,各x_iにおけるデータになるべく$g(x)$ をフィットさせようとしている.
- 第2項は粗さのペナルティ(*roughness penalty*)であり, $g(x)$の変動の程度を示す. チューニング母数(*tuning parameter*) $\lambda \geq 0$により制御する.
  - λが小さければ関数はより変動し, $\lambda = 0$の時にはy_iの補間になる.

平滑化スプライン-Smoothing Splines-

ここでの議論は少し数理的

滑らかな関数$g(x)$をあるデータにフィットする

次の規準を考える:

$$\underset{g\in\mathcal{S}}{\text{minimize}} \sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- 第1項はRSSであり,各x_iにおけるデータになるべく$g(x)$ をフィットさせようとしている.
- 第2項は粗さのペナルティ(*roughness penalty*)であり, $g(x)$の変動の程度を示す. チューニング母数(*tuning parameter*) $\lambda \geq 0$により制御する.
  - λが小さければ関数はより変動し, $\lambda = 0$の時にはy_iの補間になる.
  - $\lambda \to \infty$のとき関数$g(x)$は線形になる.

平滑化スプライン-続き

自然3次スプラインで各x_iを結節点としよう. このときでさえ粗さペナルティにより粗さは λを通じて制御される.

平滑化スプライン-続き

自然3次スプラインで各x_iを結節点としよう. このときでさえ
粗さペナルティにより粗さは λを通じて制御される.

長所

- 平滑化スプライン(Smoothing splines)により結節点の選択問題を避け, パラメター λ を選べばよい.

平滑化スプライン-続き

自然3次スプラインで各x_iを結節点としよう. このときでさえ
粗さペナルティにより粗さは λを通じて制御される.

長所

- 平滑化スプライン(Smoothing splines)により結節点の選択問題を避け, パラメター λ を選べばよい.
- アルゴリズムはここで説明するには複雑すぎる. Rの中では smooth.spline() により平滑化スプラインを実行できる.

平滑化スプライン-続き

自然3次スプラインで各x_iを結節点としよう. このときでさえ
粗さペナルティにより粗さは λを通じて制御される.

長所

- 平滑化スプライン(Smoothing splines)により結節点の選択問題を避け, パラメター λ を選べばよい.
- アルゴリズムはここで説明するには複雑すぎる. Rの中では smooth.spline() により平滑化スプラインを実行できる.
- n個のフィットした値を$\hat{\boldsymbol{g}}_{\boldsymbol{\lambda}} = \boldsymbol{S}_{\boldsymbol{\lambda}}\boldsymbol{y}$とする. ただし$\boldsymbol{S}_{\boldsymbol{\lambda}}$は行列とする($x_i$および$\lambda$により定まる).
- 有効な自由度(*effective degrees of freedom)*は次式で与えられる

$$df_\lambda = \sum_{i=1}^{n} \{\boldsymbol{S}_{\boldsymbol{\lambda}}\}_{ii}$$

平滑化スプライン-続き

- λではなくdf を特定化する!
  Rでは: smooth.spline(age, wage, df = 10)とすればよい.

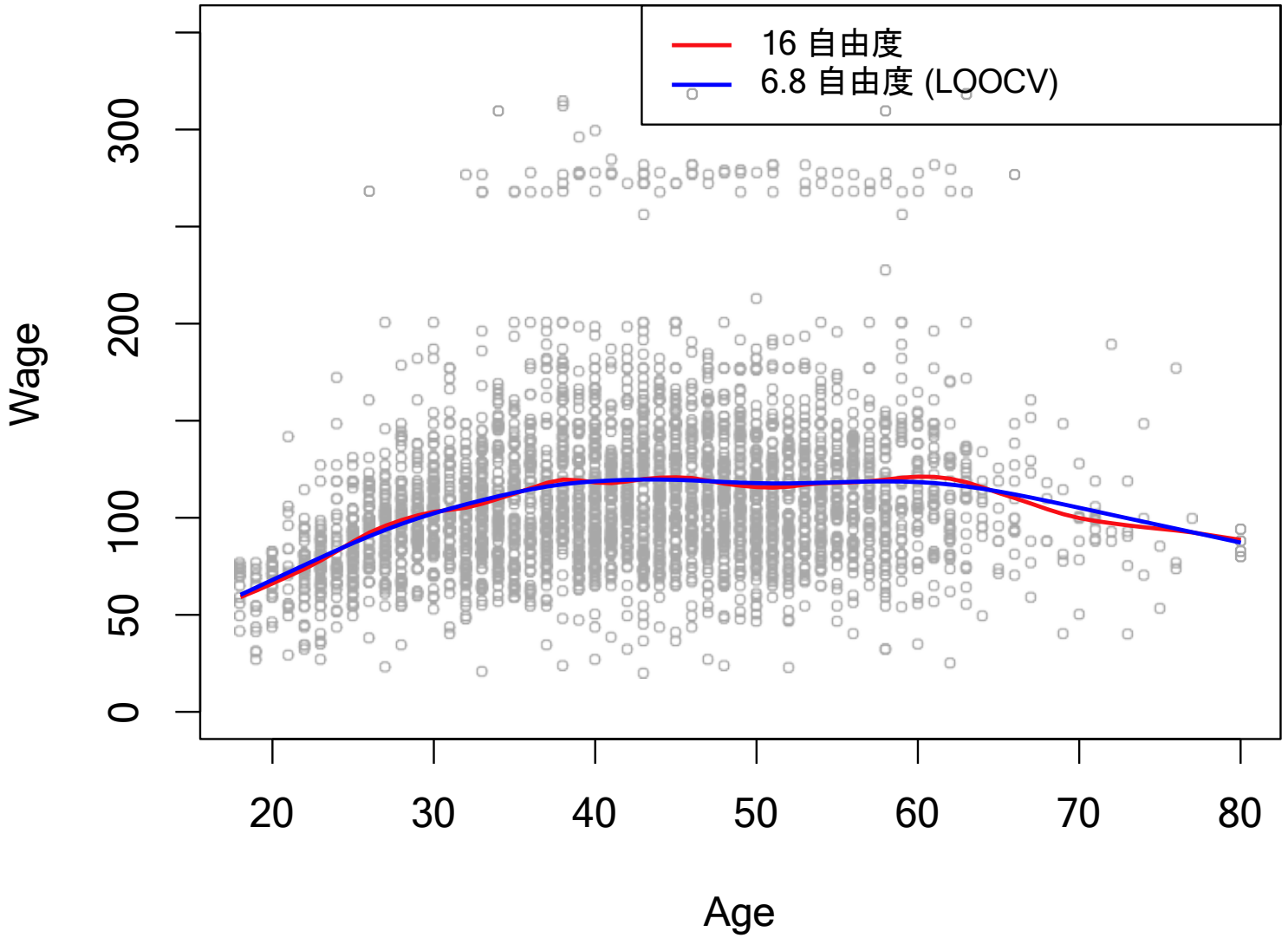
平滑化スプライン-続き

- λではなくdf を特定化する!
  Rでは: smooth.spline(age, wage, df = 10)とすればよい.
- leave-one-out (LOO) 交差検証誤差は次式で与えられる

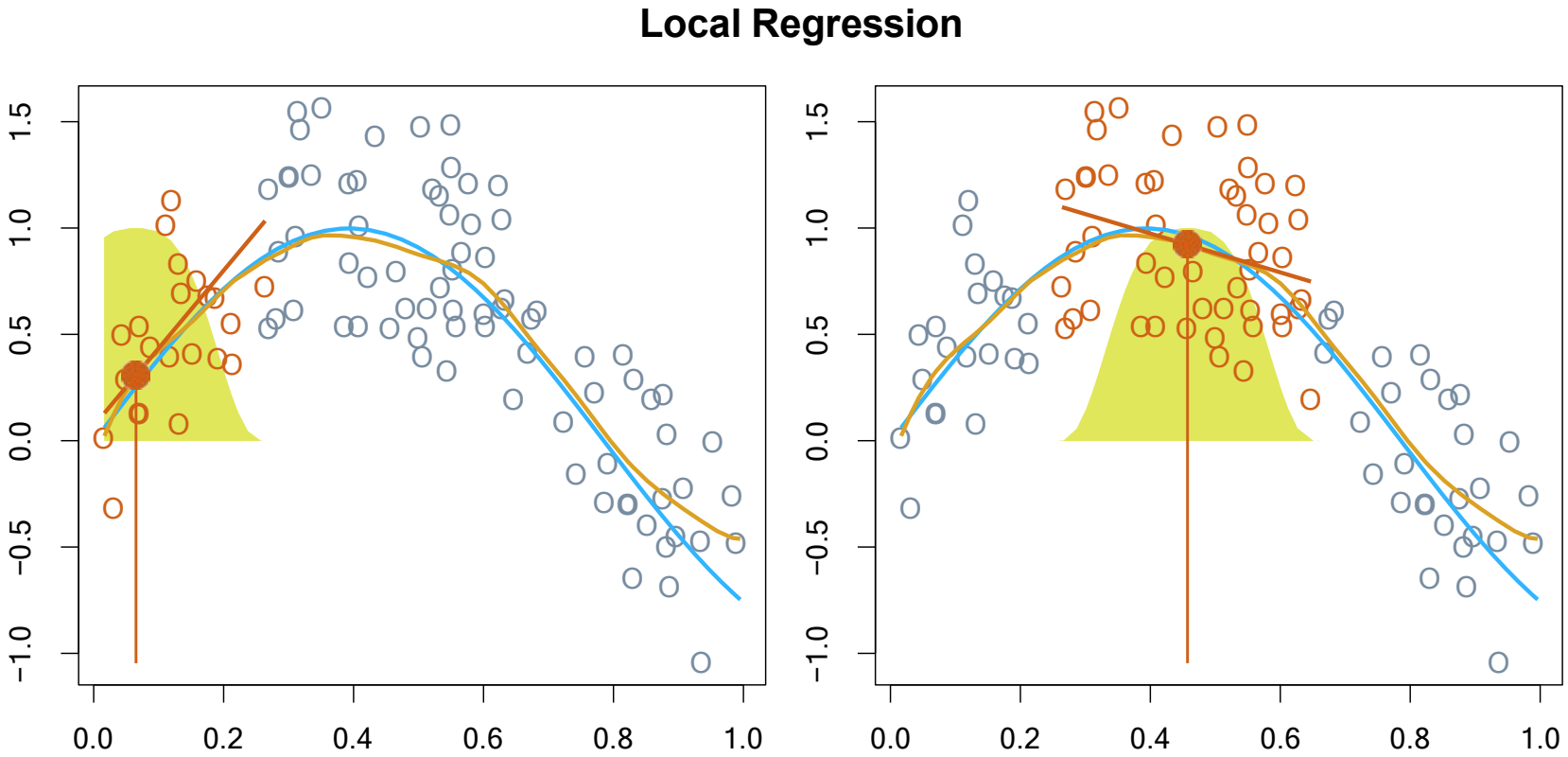
$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_{\lambda}^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - \{\boldsymbol{S}_{\boldsymbol{\lambda}}\}_{ii}} \right]^2 .$$

  Rでは: smooth.spline(age, wage)

平滑化スプライン(Smoothing Spline)



局所回帰-Local Regression-



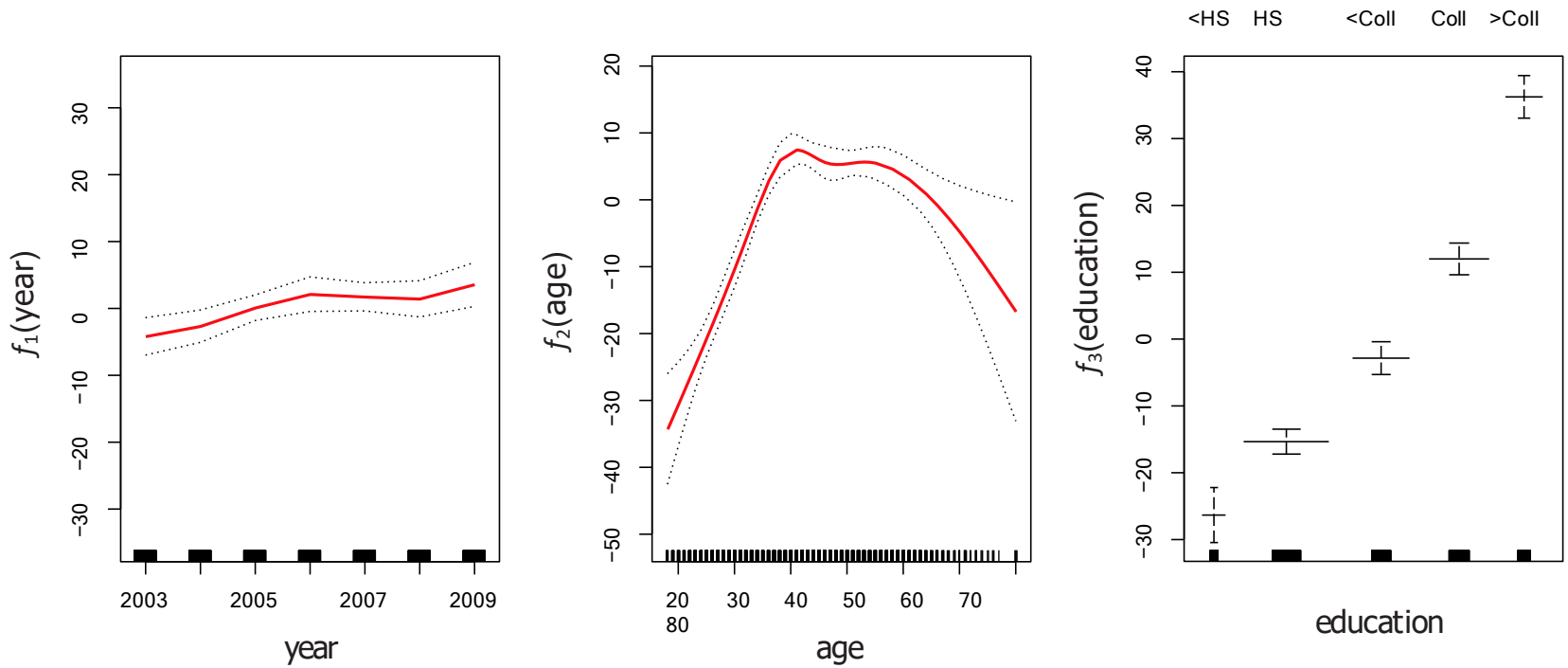
ウエイト関数を変化させつつ加重最小二乗法によりXの範囲で線形フィットを行う.

詳細はテキストを参照, Rでは loess() 関数を利用.

一般化加法モデル-Generalized Additive Models-

複数の変数についてより柔軟な非線形性を認めつつ, 線形モデルの構造を保つ.

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i.$$



一般化加法モデル(GAM)

- GAMのフィットには次を用いる, 例えば自然スプライン(natural splines)なら:
  lm(wage ~ ns(year, df = 5) + ns(age, df = 5) + education)

一般化加法モデル(GAM)

- GAMのフィットには次を用いる, 例えば自然スプライン(natural splines)なら:
  lm(wage ~ ns(year, df = 5) + ns(age, df = 5) + education)

- 係数自体はそれほど関心がなく, フィットした関数に関心がある.
  前の図はコマンド plot.gamを用いて作成した.

一般化加法モデル(GAM)

- GAMのフィットには次を用いる, 例えば自然スプライン(natural splines)なら:
  lm(wage ~ ns(year, df = 5) + ns(age, df = 5) + education)

- 係数自体はそれほど関心がなく, フィットした関数に関心がある. 前の図はコマンド plot.gamを用いて作成した.
- 項を混ぜることが可能 — 幾つかは線形, 幾つかは非線形 — とすると、モデルを比較するにはコマンドanova() を用いる.

一般化加法モデル(GAM)

- GAMのフィットには次を用いる, 例えば自然スプライン(natural splines)なら:
  lm(wage ~ ns(year, df = 5) + ns(age, df = 5) + education)

- 係数自体はそれほど関心がなく, フィットした関数に関心がある. 前の図はコマンド plot.gamを用いて作成した.
- 項を混ぜることが可能 — 幾つかは線形, 幾つかは非線形 — とすると、モデルを比較するにはコマンドanova() を用いる.
- 平滑化スプラインsmoothing splines, 局所回帰local regressionも利用できる:
  gam(wage ~ s(year, df = 5) + lo(age, span = .5) + education)

一般化加法モデル(GAM)

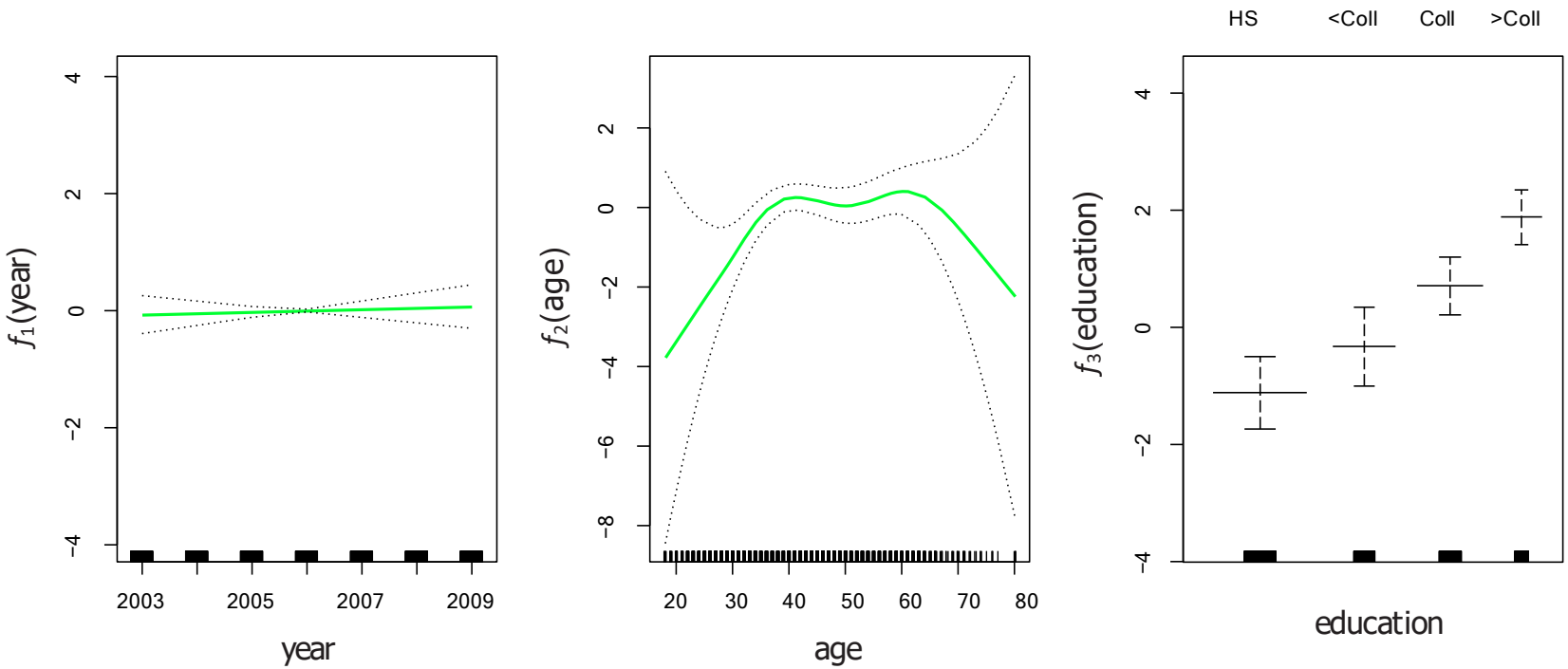
- GAMのフィットには次を用いる, 例えば自然スプライン(natural splines)なら:
  lm(wage ~ ns(year, df = 5) + ns(age, df = 5) + education)

- 係数自体はそれほど関心がなく, フィットした関数に関心がある. 前の図はコマンド plot.gamを用いて作成した.
- 項を混ぜることが可能 — 幾つかは線形, 幾つかは非線形 — とすると、モデルを比較するにはコマンドanova() を用いる.
- 平滑化スプラインsmoothing splines, 局所回帰local regressionも利用できる:
  gam(wage ~ s(year, df = 5) + lo(age, span = .5) + education)

- GAMは加法的であり, 自然に低次数の交互作用を含めることもでき, 例えば2次平滑化や次の形の交互作用
  ns(age,df=5):ns(year,df=5).

分類のためのGAM

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$



gam(I(wage > 250) ~ year + s(age, df = 5) + education, family = binomial)

第8章 木に基づく方法
-Tree-based methods-

- 決定木とは
- 回帰木
- 木の剪定
- 分類木
- ジニ指数と交差エントロピー
- バギング
- ランダムフォレスト
- ブースティング
- 変数の重要度

第8章 木に基づく方法
-Tree-based methods-

- 回帰と分類のための木に基づく方法について説明する。
- これらの方法においては、予測変数空間をいくつかの単純な領域に分割または層別化する。
- 予測変数空間を分割するために使用される分割ルールは木構造を用いてまとめることができるため、これらの手法は決定木手法として知られている。

長所と短所

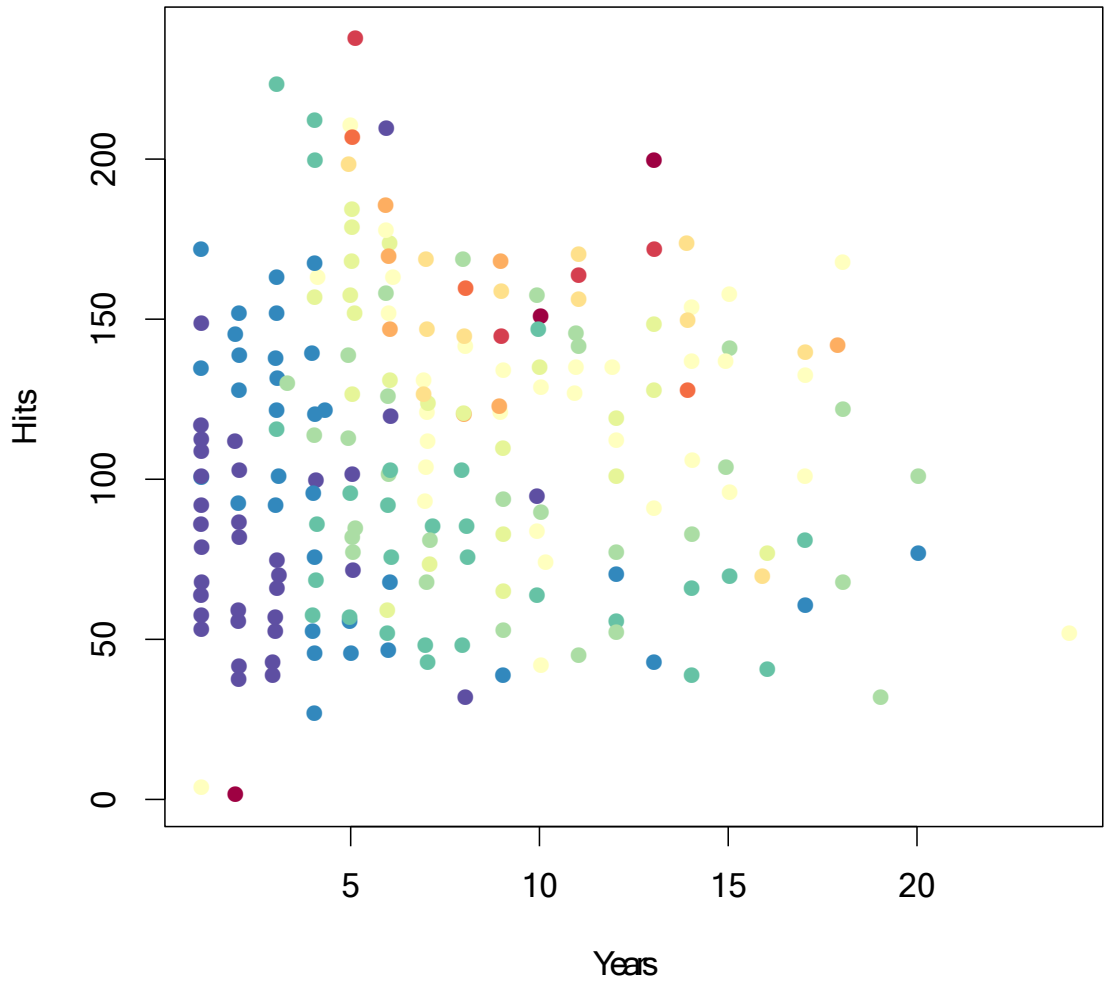
- 木に基づく方法はシンプルで解釈が容易である。
- しかしその予測精度は、最良の教師あり学習アプローチに劣る。
- したがって、バギング、ランダムフォレスト、ブースティングについても説明する。これらの手法は複数の木を作成し、1つの合意予測を生成するために組み合わされる。
- 多数の木を組み合わせることで、予測精度を劇的に向上させることができるが、解釈のしやすさについて若干の犠牲を払う。

決定木の基本

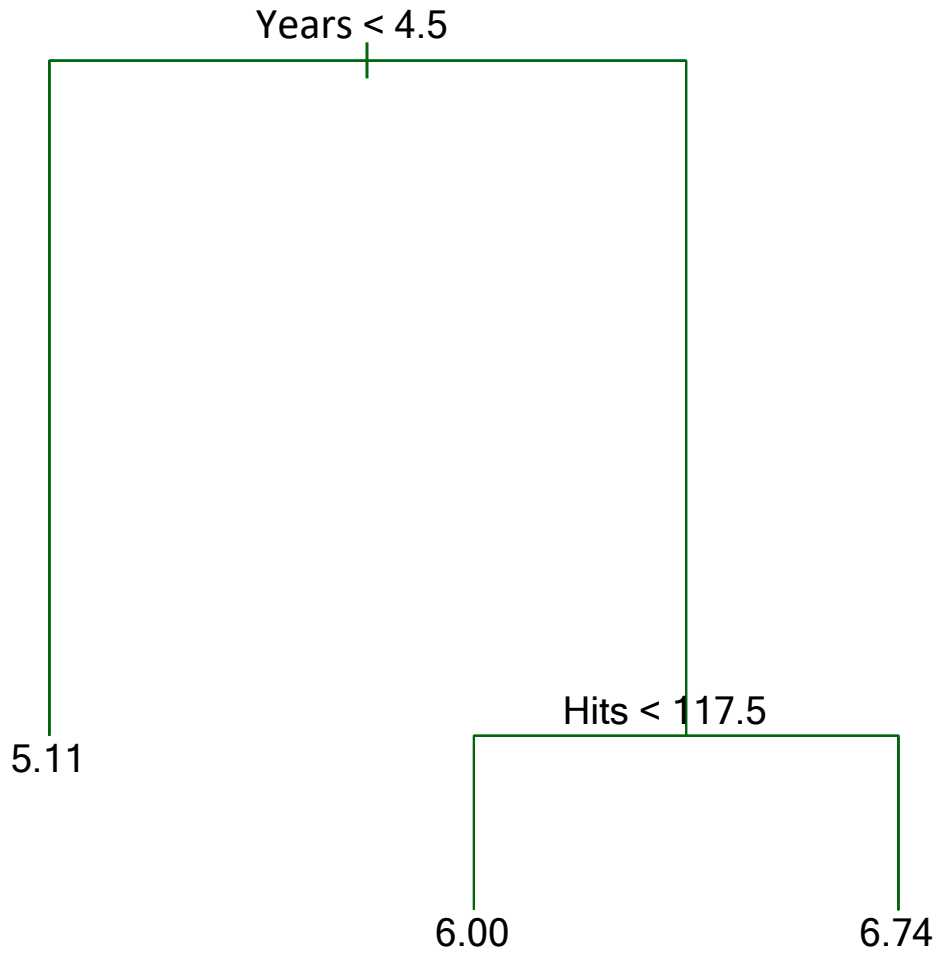
- 決定木は回帰問題と分類問題の両方に適用できる。
- まず回帰問題を考え、次に分類問題に移る。

野球選手の収入データ:どのように層別化しますか？

収入は、低いもの（青、緑）から高いもの（黄色、赤）に色分けされている。



野球選手の収入データ:決定木

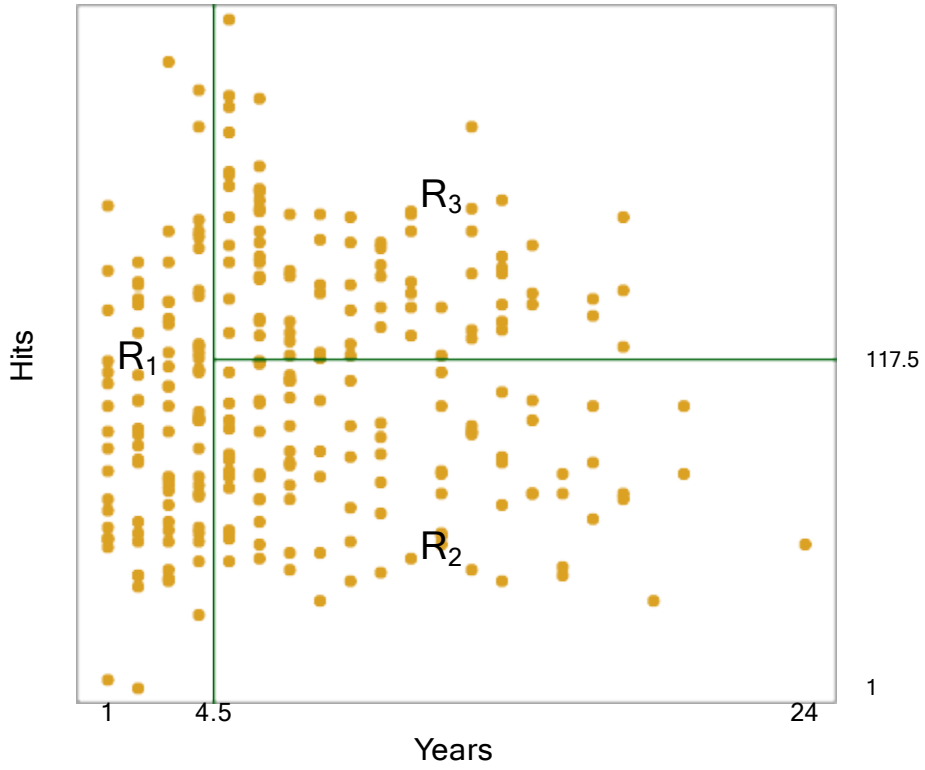


前の図について

- Hittersデータセットに関して、野球選手のメジャーリーグ在籍年数と前年度のヒット数に基づいて、対数変換した収入を予測するための回帰木である。
- それぞれの内部ノードでは、ラベル（$X_j < t_k$の形式）は、その分割から生じる左側の枝を示し、右側の枝は$X_j \geq t_k$に対応する。たとえば、頂点における分割は2つの大きな枝になる。左側の枝はYears < 4.5に対応し、右側の枝はYears≥4.5に対応する。
- 木には2つの内部ノードと3つの終端ノードまたは葉がある。各葉の数字は、そこに該当する観測値の応答変数の平均である。

結果

- 全体的に、この木は予測空間を3つの領域に分割している：R_1 = {X | Years< 4.5}, R_2 = {X | Years>=4.5, Hits<117.5}, R_3 = {X | Years>=4.5, Hits>=117.5}



木に関する用語

- 木に例えていることから、R1、R2、R3の領域は葉ノードと呼ばれる。
- 決定木は通常、葉が木の下部にあるという意味で逆さまに描かれる。
- 予測変数空間が分割される木の点を内部ノードと呼ぶ。
- Hitters treeにおいては、２個の内部ノードがYears<4.5とHits<117.5として示されている。

結果の解釈

- 収入(Salary)を決定する上で最も重要な要素は年数(Years)であり、経験の浅い選手ほど収入が低くなる傾向にある。
- 選手が経験不足である場合、前年にヒットをした数(Hits)は収入にあまり影響しない。
- しかし、5年以上メジャーリーグに在籍している選手の中では、前年にヒットをした数が収入に影響し、昨年ヒット数が多かった選手ほど高い収入を得ている。
- 単純化されているかもしれないが、回帰モデルと比較して、解釈や説明が容易であることが特徴である。

決定木を構成するためのプロセス

- 予測変数空間（つまり、$X_1, X_2, \ldots, X_p$のとりうる値の集合）を、J個の異なる非重複領域$R_1, R_2, \ldots, R_J$に分割します。
- 領域R_jに属する観測値ごとに、単純にR_j内の訓練データにおける応答変数の平均である同じ予測を行う。

決定木を構成するためのプロセス

- 理論上、領域は任意の形状を取ることができるが、結果として得られる予測モデルの解釈を容易にするために、高次元の長方形またはボックスに予測空間を分割することを選択する。
- 目標は、RSSを最小化するボックス$R_1, R_2, \ldots, R_J$を見つけることである。RSSは以下の式で与えられる。

$$\sum_{j=1}^{J} \sum_{j \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

- ここで、$\hat{y}_{R_j}$は、j番目のボックスに属する訓練データの応答変数の平均を表す。

決定木を構成するためのプロセス

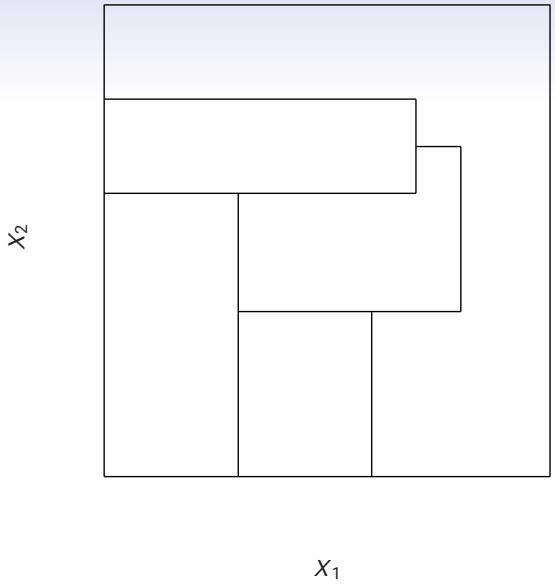
- 残念ながら、特徴空間をJ個の箱に分割する可能性すべてを検討することは計算上不可能である。
- そのため、再帰的な2分割として知られるトップダウンで貪欲なアプローチを採用している。
- このアプローチは、木の頂点から始まり、予測変数空間を連続的に分割するため、トップダウン型である。各分割は、木のさらに下の 2 つの新しい枝で示される。
- 貪欲とは、木を構成する各段階において、将来的に良い分割を選ぶのではなく、その段階で最適な分割を行うことである。

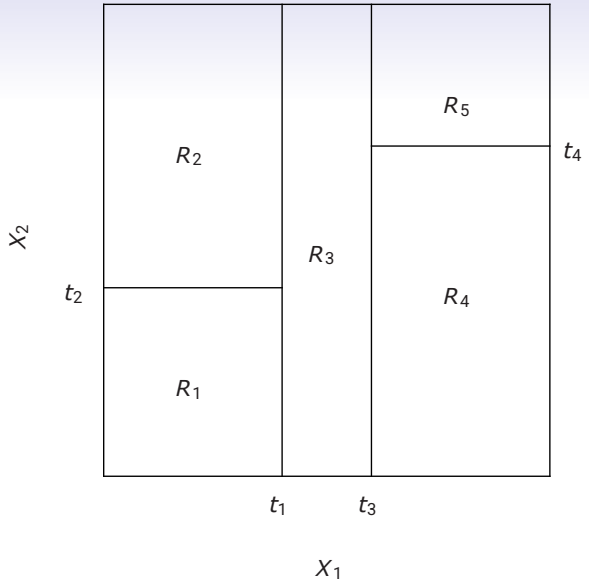
続き

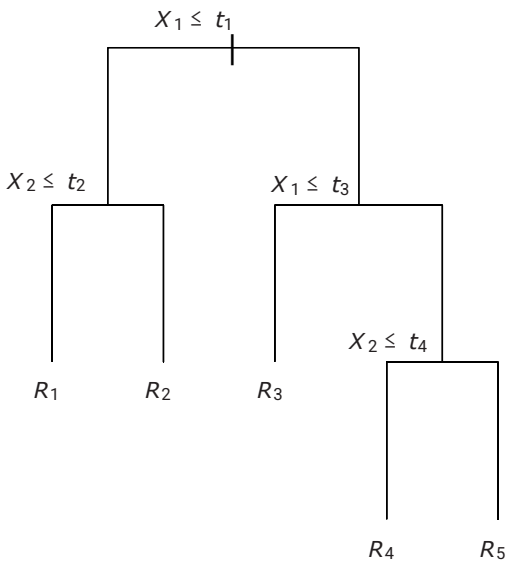
- RSSを最大限に減らすように、最初に予測変数X_jとカットポイントsを選択して、予測変数空間を$\{X|X_j < s\}$と$\{X|X_j \geq s\}$の領域に分割する。
- 次に、このプロセスを繰り返し、得られた各領域内のRSSを最小化するようにデータを更に分割するために、最適な予測変数と最適なカットポイントを探す。
- しかし今回は、予測空間全体を分割するのではなく、以前に識別された2つの領域のうちの1つを分割した。これで、3つの領域ができた。
- 再び、RSSを最小化するように、3 つの領域の1つをさらに分割する。この過程は、停止基準に達するまで続けられる。たとえば、どの領域も 5 つ以上の観測値を含まないまで続けることができる。

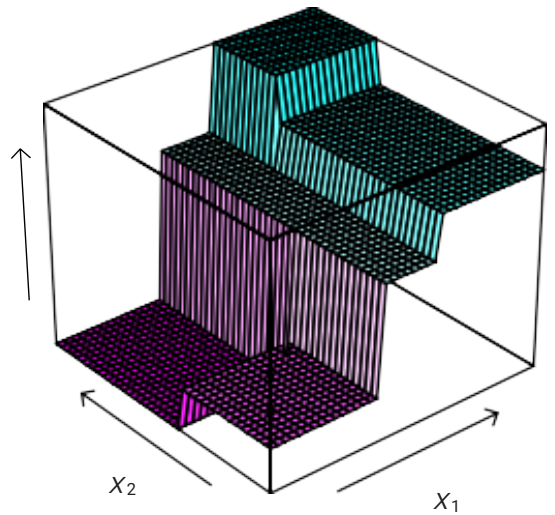
予測

- 与えられたテスト観測の応答を、そのテスト観測が属する領域内の訓練観測の平均を用いて予測する。
- このアプローチの5つの領域の例が次のスライドに示されている。


X_2
X_1


R_5
R_2
t_4
X_2
R_3
t_2
R_4
R_1
t_1
t_3
X_1


$X_1 \leq t_1$
$X_2 \leq t_2$
$X_1 \leq t_3$
$X_2 \leq t_4$
R_1
R_2
R_3
R_4
R_5


X_2
X_1

前の図の詳細

左上：再帰的二分割では得られない 2 次元特徴空間の分割。
右上：2 次元の例に対する再帰的二分割による結果。
左下：右上の分割に対応する木。
右下：その木に対応する予測平面を透視投影したもの。

木の剪定

- 上記のプロセスでは、訓練データでは良い予測ができるかもしれないが、データを過学習させてしまい、テストデータのパフォーマンスが低下してしまう可能性がある。

木の剪定

- 上記のプロセスでは、訓練データでは良い予測ができるかもしれないが、データを過学習させてしまい、テストデータのパフォーマンスが低下してしまう可能性がある。
- 分割数が少ない（つまり、領域$R_1, R_2, \ldots, R_J$の数が少ない）小さな木は、多少のバイアスを犠牲にして、分散を低くし、解釈を良くすることができる。
- 上記のプロセスの代替案は、各分割によるRSSの減少がなんらかの（高い）閾値を超える間だけ、木を成長させることである。
- この戦略は木を小さくするが、近視眼的すぎる。一見価値のないように見える分割が、後に非常に良い分割（RSSを大きく減らすことにつながる）が続くかもしれない。

木の剪定–続き

- より良い戦略は、非常に大きな木T_0を成長させ、次に部分木を得るために切り戻す（剪定する）。
- 木の複雑さをコストとした剪定法（Cost complexity pruning）は、最弱リング剪定法（weakest link pruning）とも呼ばれ、よく使われる方法である。
- すべての部分木を考えるのではなく、非負のチューニングパラメータαを変化させて得られる木の列を考える。αの各値に対して

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T|$$

- が可能な限り小さくするような部分木 $T \subset T_0$ が対応する。ここで、$|T|$はTの葉ノード数、R_mはm番目の葉ノードに対応する矩形（すなわち、予測空間の部分集合）、$\hat{y}_{R_m}$はR_mにおける応答変数の予測値、すなわちR_mの訓練データの平均である。

最適な部分木の選択

- チューニングパラメータα は、部分木の複雑さと訓練データへの当てはまりの良さとのトレードオフを制御する。
- 交差検証により、最適値$\hat{\alpha}$を選択できる。
- その後、全データセットに戻り、$\hat{\alpha}$に対応する部分木を得る。

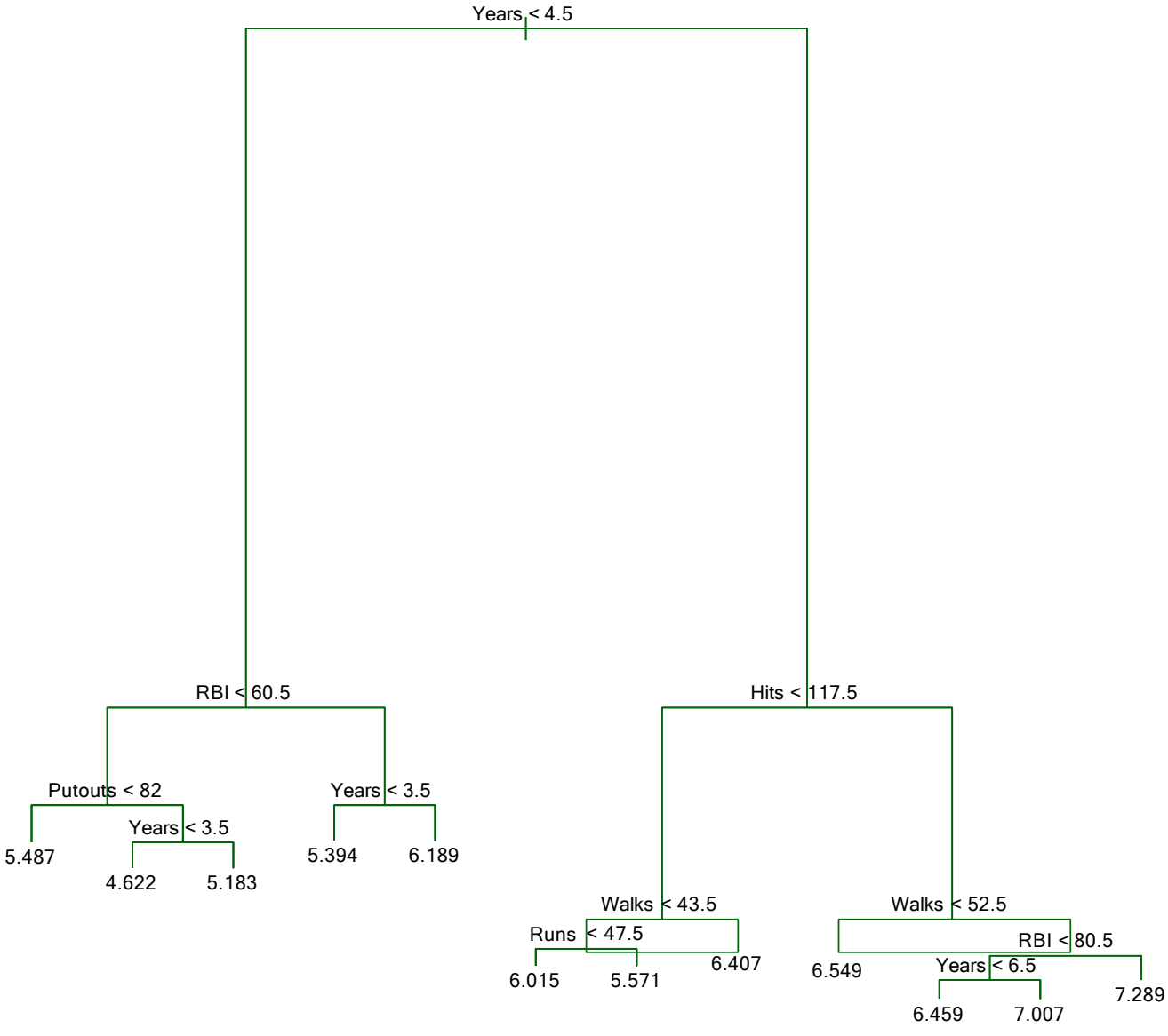
回帰木のアルゴリズム

1. 再帰的二分割を使用して、訓練データ上で大きな木を成長させ、各葉ノードがある最小数の観測より少ないときだけ停止する。
2. 最弱リング剪定法を大木に適用し、α の関数として、最適な部分木の列を得る。
3. K-fold 交差検証で α を選択する。各$k = 1, \ldots, K$に対して
    a. k 番目のデータを除いて、訓練データの $\frac{K-1}{K}$ 番目のデータでステップ 1と 2 を繰り返す
    b. α の関数として、除外された k 番目のデータのデータにおけるMSEを評価する。各αについてこれらの結果を平均し、平均誤差を最小化するαを選ぶ。
4. ステップ2にある部分木の列から、選択したαに対応する部分木を出力する。

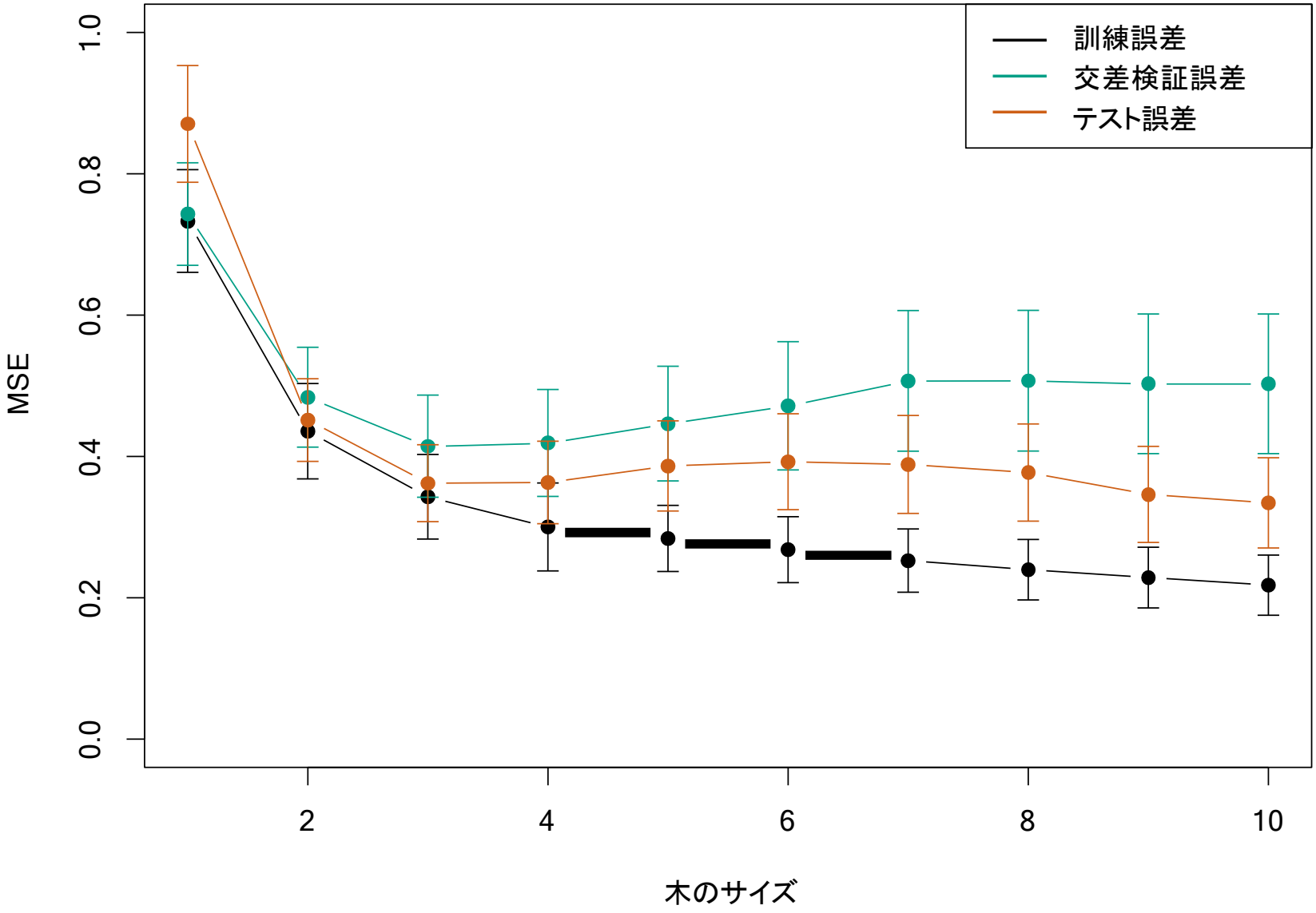
野球の例の続き

- 最初に、データセットをランダムに2分割し、訓練データに 132 の観測値、テストデータに 131 の観測値を得た。
- 次に、訓練データに対して大きな回帰木を作り、α を変化させることで、異なる葉ノード数を持つ部分木を作成した。
- 最後に，6 分割交差検証を行い、α の関数として木の交差検証 MSE を推定した。

野球の例の続き


Years < 4.5
RBI < 60.5
Hits < 117.5
Putouts < 82
Years < 3.5
Years < 3.5
Walks < 43.5
Walks < 52.5
Runs < 47.5
RBI < 80.5
Years < 6.5
5.487
4.622
5.183
5.394
6.189
6.015
5.571
6.407
6.549
6.459
7.007
7.289

野球の例の続き



分類木

- 回帰木とよく似ているが、量的ではなく質的な応答変数を予測するために使用される。
- 分類木の場合、各観測値は、それが属する領域において、訓練データの最も頻度の高いクラスに属すると予測する。

分類木の詳細

- 回帰の設定と同様に、再帰的二分割を用いて分類木を成長させる。
- 分類の設定では、RSS は 2 値分割を行うための基準として使用することはできない。
- RSS の自然な代替は、誤分類率である。これは単に、最も一般的なクラスに属さないその領域での訓練データの割合である。

$$E = 1 - \max_k(\hat{p}_{mk})$$

- ここで$\hat{p}_{mk}$は、m 番目の領域における訓練データがk 番目のクラスからのものである比率を表す。
- ただし、誤分類率は、木の成長に対して十分な感度を持たず、実際には、他の 2 つの尺度が望ましい。

ジニ指数とデビアンス

- ジニ指数は

$$G_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

訳注：原文ではGとなっているが、G_mに修正している。

によって定義され、K個のクラス間の分散を表す尺度である。ジニ指数は、すべての$\hat{p}_{mk}$の値が0または1に近い場合に小さい値をとる。

- この理由から、ジニ指数はノードの純度に対する尺度として用いられる。つまり、小さい値であることは、ほとんどの観測値が同じクラスに属しているノードであることを表す。
- ジニ指数の代替として、交差エントロピーがある。

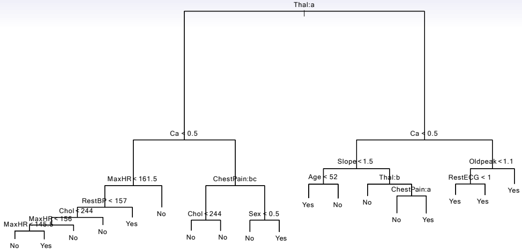
$$D_m = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

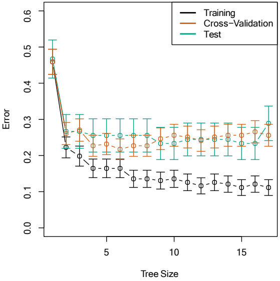
訳注：原文のDをD_mと修正している。

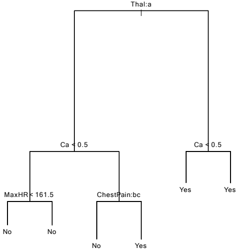
- 実際、ジニ指数と交差エントロピーは数値的に非常に似ていることがわかる。

心臓病データの例

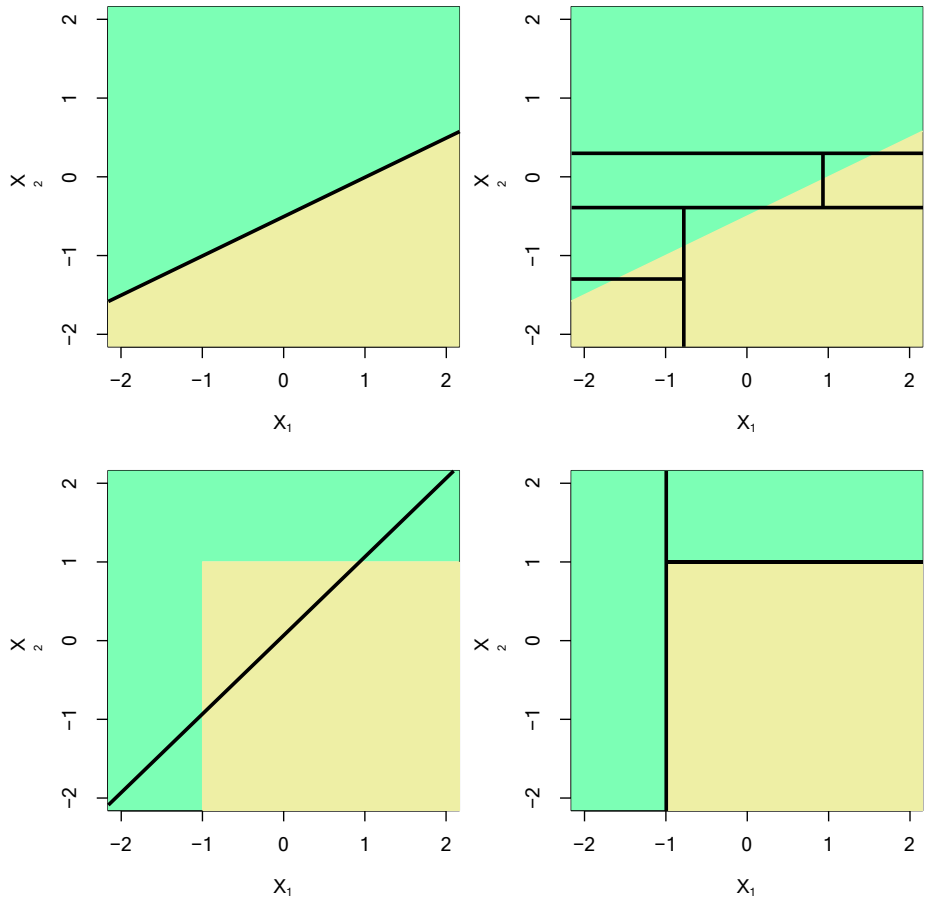
- これらのデータには、胸痛を訴えた303人の患者における2値変数「HD」が含まれている。
- Yesの場合は、血管造影検査に基づく心臓病の存在を示し、Noは心臓病がないことを示す。
- 年齢(Age)、性別(Sex)、コレステロール測定値(Chol)、他の心臓および肺機能の測定値を含めた13個の予測変数がある。
- 交差検証により、6つの葉ノードを持つ木が生成される。次の図を参照してください。


Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
Slope < 1.5
Oldpeak < 1.1
Age < 52
Thal:b
ChestPain:a
RestECG < 1
RestBP < 157
Chol < 244
MaxHR < 156
MaxHR < 145.5
Chol < 244
Sex < 0.5
No
Yes


Training
Cross−Validation
Test
Error
Tree Size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
5
10
15


Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
No
Yes

木と線形モデル



上段：真の線形境界；下段：真の非線形境界。
左列：線形モデル；右列：木に基づくモデル

木の利点と欠点

- 木は人々に説明するのが非常に簡単である。実際、線形回帰よりも説明しやすい！
- 一部の人々は、回帰や分類手法よりも決定木が人間の意思決定に近く、より類似していると考えている。
- 木はグラフィカルに表示でき、小さい木であれば非専門家でも簡単に解釈できる。
- 木はダミー変数を作成する必要がないため、質的的な予測変数を簡単に処理できる。
- しかし、木は一般に、この本に記載されている他の回帰や分類手法と同じレベルの予測精度を持っていない。
- ただし、多数の決定木を集約することにより、木の予測性能を大幅に改善できる。これらの概念を次に紹介する。

バギング

- ブートストラップ集約法、またはバギングは、統計的学習方法の分散を低減するための汎用的な手法であり、ここで導入するのは、特に決定木の文脈で特に有用で頻繁に使用されるためである。
- 独立したn個の観測値$Z_1, \ldots, Z_n$が与えられ、それぞれが分散σ^2を持つ場合、観測値の平均$\bar{Z}$の分散はσ^2/nで与えられる。
- 観測値の平均を取ることで分散が低減されることを思い出してください。もちろん、これは実際的ではない。なぜなら、通常は複数の訓練データセットを得ることはないからである。

バギング-続き

- 代わりに、訓練データから繰り返し標本を抽出することによってブートストラップができる。
- このアプローチでは、B個の異なるブートストラップ訓練データを生成する。その後、私たちはb番目のブートストラップ訓練データで私たちの方法を学習して、点xでの予測$\hat{f}^{*b}(x)$を得る。その後、すべての予測値を平均して、

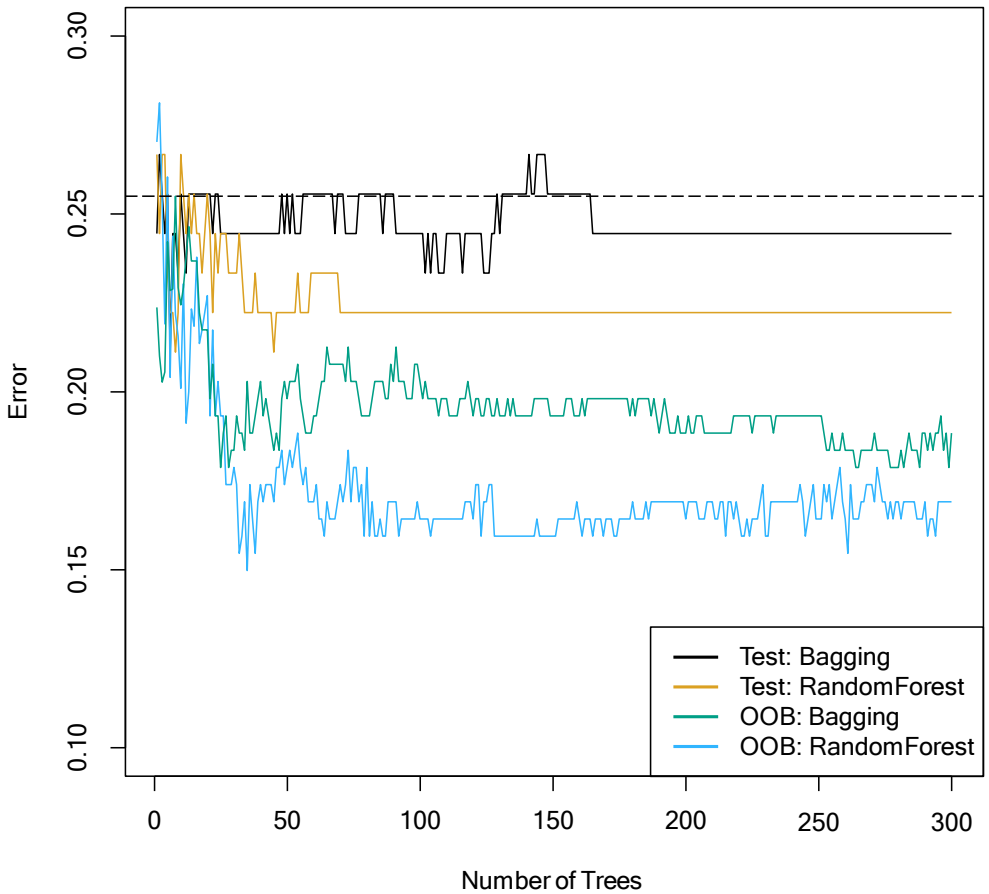
$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$

- を得る。これをバギングと呼ぶ。

バギング分類木

- 上記の方法は回帰木に適用される。
- 分類木の場合：各テストデータについて、B本の木それぞれが予測したクラスを記録し、多数決を行う。つまり、B個の予測のうち最も一般的なクラスが全体的な予測となる。

心臓病データにおけるバギングの結果



前の図について

心臓病データに対するバギングとランダムフォレストの結果

- テスト誤分類率（黒とオレンジ）が、訓練データセットからブートストラップにより作成した木の数Bの関数として表されている。
- ランダムフォレストは、$m = \sqrt{p}$で適用された。
- 点線は、単一の分類木から得られるテスト誤分類率を示している。
- 緑と青の線は、この場合はかなり低いOOB誤分類率を示している。

Out-of-bagによる誤分類率

- バギングモデルのテスト誤分類を推定するのは非常に簡単である。
- バギングの鍵は、観測値のブートストラップ標本に対し木の当てはめを繰り返して行うことである。平均的に、バッグされた各木は観測値の約三分の二を使用する。
- バッグされた木に用いられなかった残りの三分の一の観測値は、out-of-bag(OOB) 観測値と呼ばれる。
- i番目の観測値の応答変数は、その観測値がOOBであった木を使って予測できる。これにより、i番目の観測値に対して$B/3$の予測値が得られ、これを平均化する。
- Bが大きい場合、この推定は実質的にバギングのLOOCV誤差である。

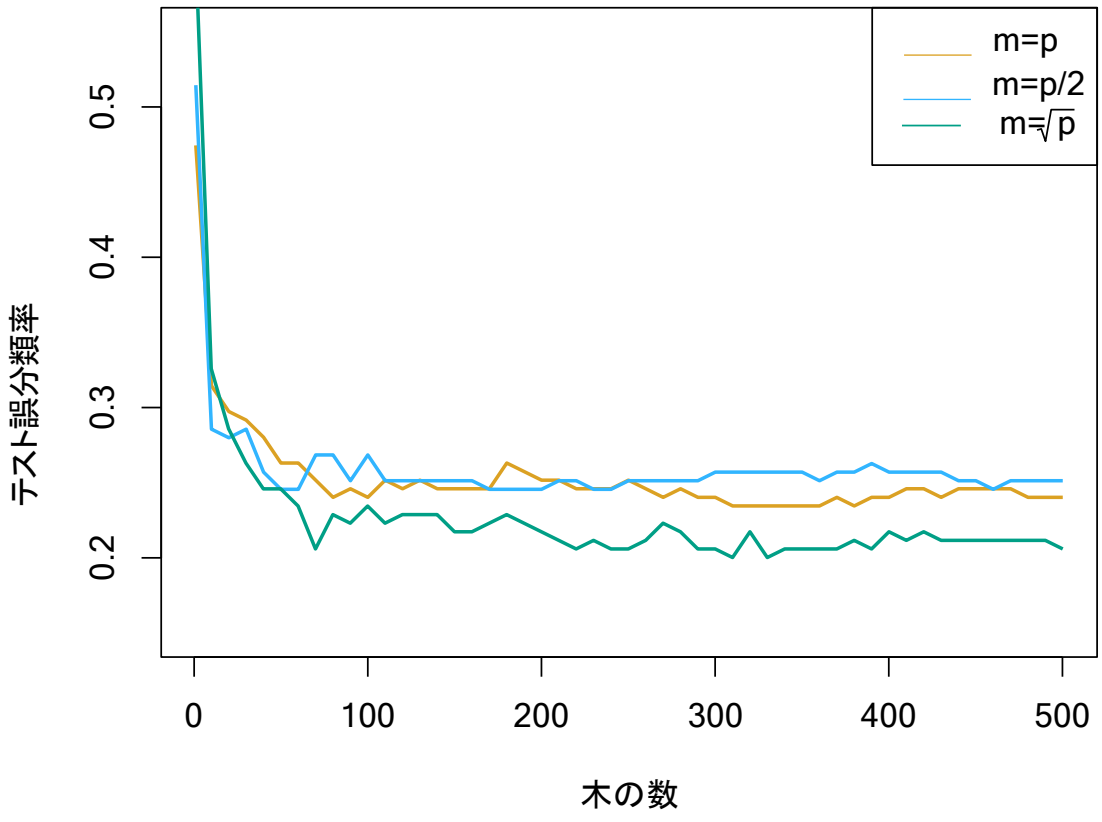
ランダムフォレスト

- ランダムフォレストは、木の相関関係をなくすような僅かな調整によって、バギングで得られた木を改良するものである。
- バギングと同様に、ブートストラップで得られた訓練データを用いていくつかの決定木を構築する。
- ただし、これらの決定木を構築するときに、木を分割するたびに、p個の全予測変数から分割の候補として、m個のランダムサンプルが選ばれる。分割では、そのm個の予測変数のうちの1つしか使用できない。
- 分割する度に新たにm個の予測変数サンプルが選ばれ、通常mの値として$m \approx \sqrt{p}$が用いられる。すなわち、各分割において考慮する予測変数の数はおおよそ全予測変数の数の平方根にほぼ等しい（心臓病データの場合は13のうち4）。

遺伝子発見データ

- 私たちはランダムフォレストを用いて、349人の組織サンプルから測定された、4,718遺伝子の発現測定から成る高次元の生物学的データセットに適用しました。
- 人間には約20,000の遺伝子があり、個々の遺伝子は、特定の細胞、組織、および生物学的条件において、異なる活性度、または発現度を持っている。
- 各患者サンプルには、15種類の定性的なラベルがある。正常か、14種類の異なるがんのいずれかである。
- 私たちは、訓練データで最も大きな分散を持つ500の遺伝子に基づいてがんのタイプを予測するためにランダムフォレストを使用している。
- 観測値を訓練データとテストデータにランダムに分割し、分割の際考慮する変数の数として3つの異なるmを用いて訓練データにランダムフォレストを適用した。

結果：遺伝子発見データ



前図の詳細

- $p = 500$の予測変数を持つ15クラスの遺伝子発現データセットに対するランダムフォレストの結果。
- テスト誤分類率は木の数の関数として表示される。各色の線は、各木の内部ノードの分割において使用可能な予測変数の数であるmの異なる値に対応している。
- ランダムフォレスト$(m < p)$はバギング$(m = p)$よりわずかに改善される。単一の分類木の誤り率は45.7%である。

ブースティング

- ブースティングは、回帰または分類のための多くの統計学習手法に適用できる一般的なアプローチであるが、ここでは、決定木におけるブースティングに限定して議論する。
- バギングでは、ブートストラップを使用して元の訓練データの複数のコピーを作成し、それぞれのコピーに別々の決定木を適合させ、すべての木を組み合わせて単一の予測モデルを作成する。
- 注目すべきことに、各木は他の木とは独立したブートストラップデータセット上に構築される。
- ブースティングはこれと同様の方法であるが、木が順次段階的に成長していく点が異なる。各木は、以前に成長した木からの情報を使用して成長する。

回帰木におけるブースティングのアルゴリズム

1. $\hat{f}(x) = 0$とし、訓練データセットのすべてのiにおいて$r_i = y_i$と設定する
2. $b = 1,2, \ldots, B$において、以下の手順を繰り返す。
    1. 訓練データ(X, r)に対し、d個の分割（$d + 1$個の葉ノード）を持つ木$\hat{f}^b$を当てはめる。
    2. 新しい木を縮小した木を加えることにより、$\hat{f}$を更新する:
    $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$$
    3. 残差を更新する:
    $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$$
3. ブースティングされたモデルを出力する。
$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$$

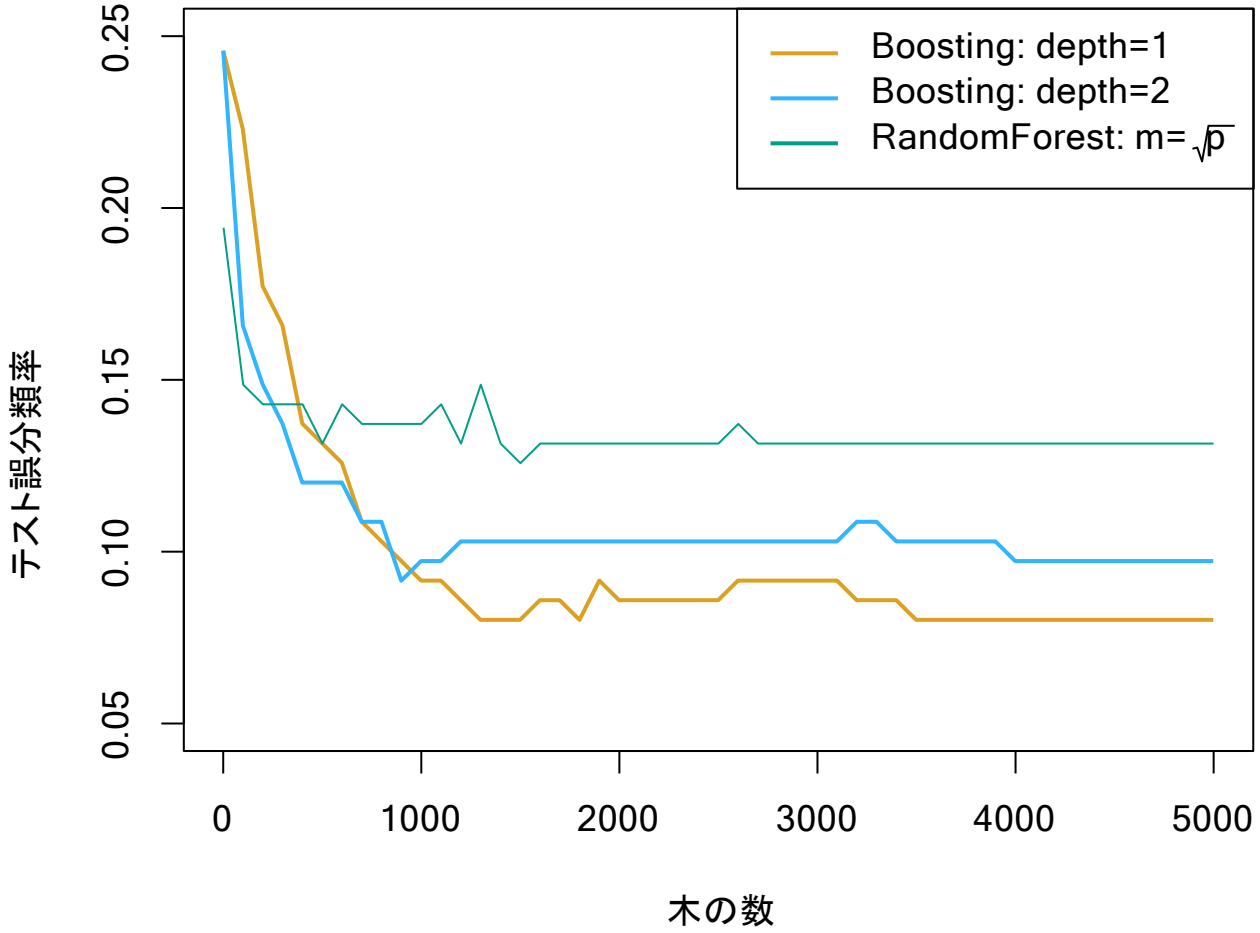
このアルゴリズムのアイディアは何ですか？

- データに単一の大きな決定木を当てはめることとは異なり、過学習を防ぐために、ブースティング手法はゆっくりと学習する。
- 現在のモデルが与えられた場合、その残差に対して決定木を当てはめる。その後、当てはめた関数にこの新しい決定木を追加して残差を更新する。
- これらの木は、アルゴリズムのパラメータdによって決定される、わずかな葉ノードのみで構成される。
- 残差に小さな木を当てはめることで、性能が悪い領域で$\hat{f}$をゆっくりと改善する。縮小パラメータλはその過程をさらに遅らせ、さらに多様な形状の木が残差に当てはめられるようにする。

分類のためのブースティング

- 分類木のブースティングも同様に構成されるが少々複雑であるので、ここでは詳細を省く。
- 学生の方はElements of Statistical Learning, chapter 10を参照されたい
- R package gbm (gradient boosted models)は多くの回帰木問題や分類問題を扱っている

遺伝子発見データ-続き


Boosting: depth=1
Boosting: depth=2
RandomForest: m= √p
テスト誤分類率
0.05
0.10
0.15
0.20
0.25
0
1000
2000
3000
4000
5000
木の数

前の図の詳細

- がんと正常の予測を行うために、ブースティングとランダムフォレストを実行した15クラスの遺伝子発現データセットの結果。
- テスト誤分類率は、木の数の関数として表示される。2つのブースティングしたモデルの場合、$\lambda = 0.01$。深さ1の木は、深さ2の木よりもわずかに優れており、どちらもランダムフォレストよりも優れているが、標準誤差は0.02程度であり、これらの差には有意性がない。
- 単一の木のテスト誤差率は24%である。

ブースティングのチューニングパラメータ

1. 木の数B。バギングやランダムフォレストとは異なり、Bが大きすぎると過学習が発生することがあるが、過学習が発生する場合でも、それは徐々に発生する傾向がある。Bの選択には交差検証を使用する。

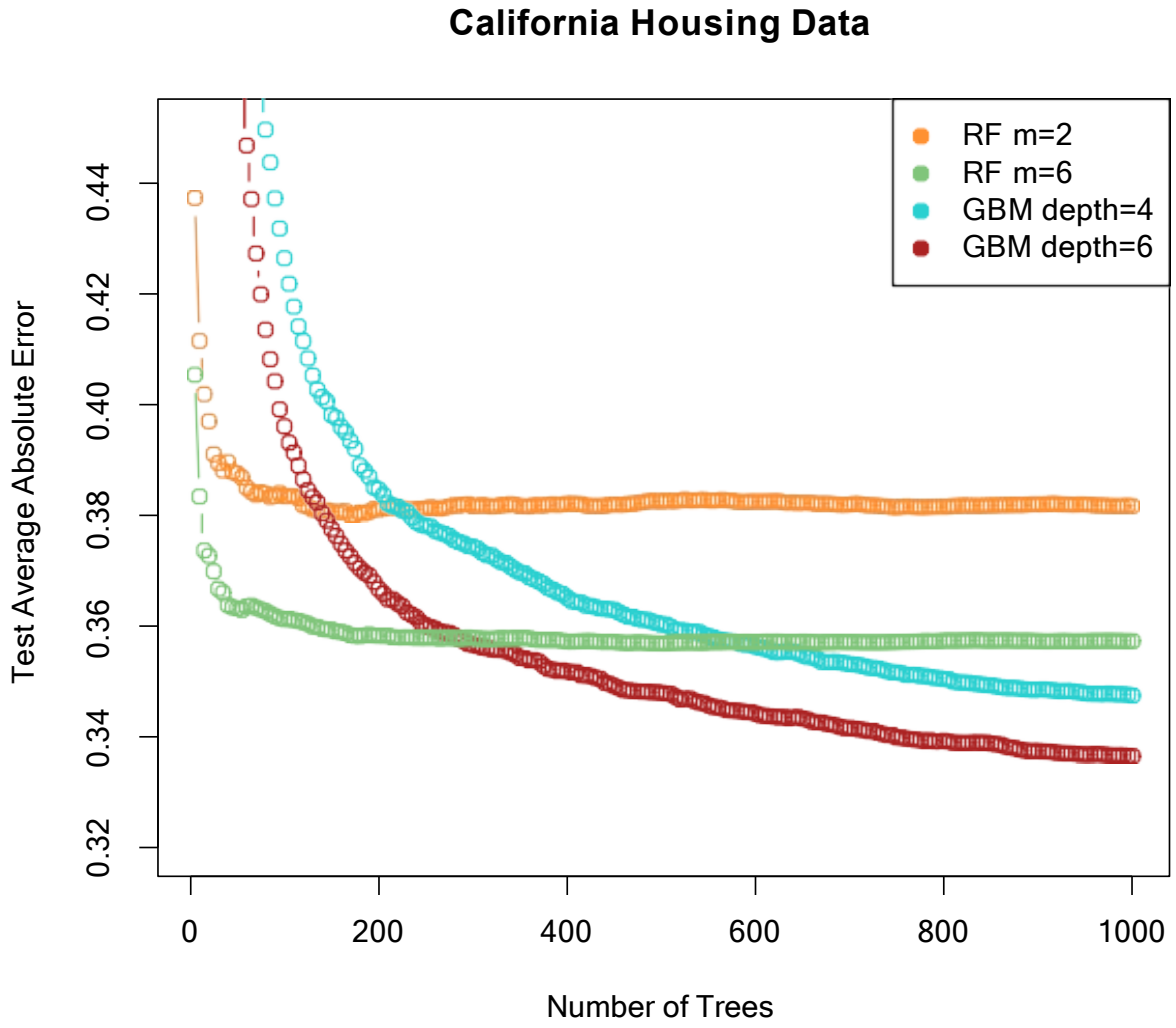
ブースティングのチューニングパラメータ

1. 木の数B。バギングやランダムフォレストとは異なり、Bが大きすぎると過学習が発生することがあるが、過学習が発生する場合でも、それは徐々に発生する傾向がある。Bの選択には交差検証を使用する。
2. 縮小パラメータλ、小さな正の数。これは、ブースティングが学習する速度を制御する。典型的な値は0.01または0.001であり、適切な選択は問題によって異なる。λが非常に小さい場合、良いパフォーマンスを達成するには非常に大きなBの値を使用する必要がある場合がある。

ブースティングのチューニングパラメータ

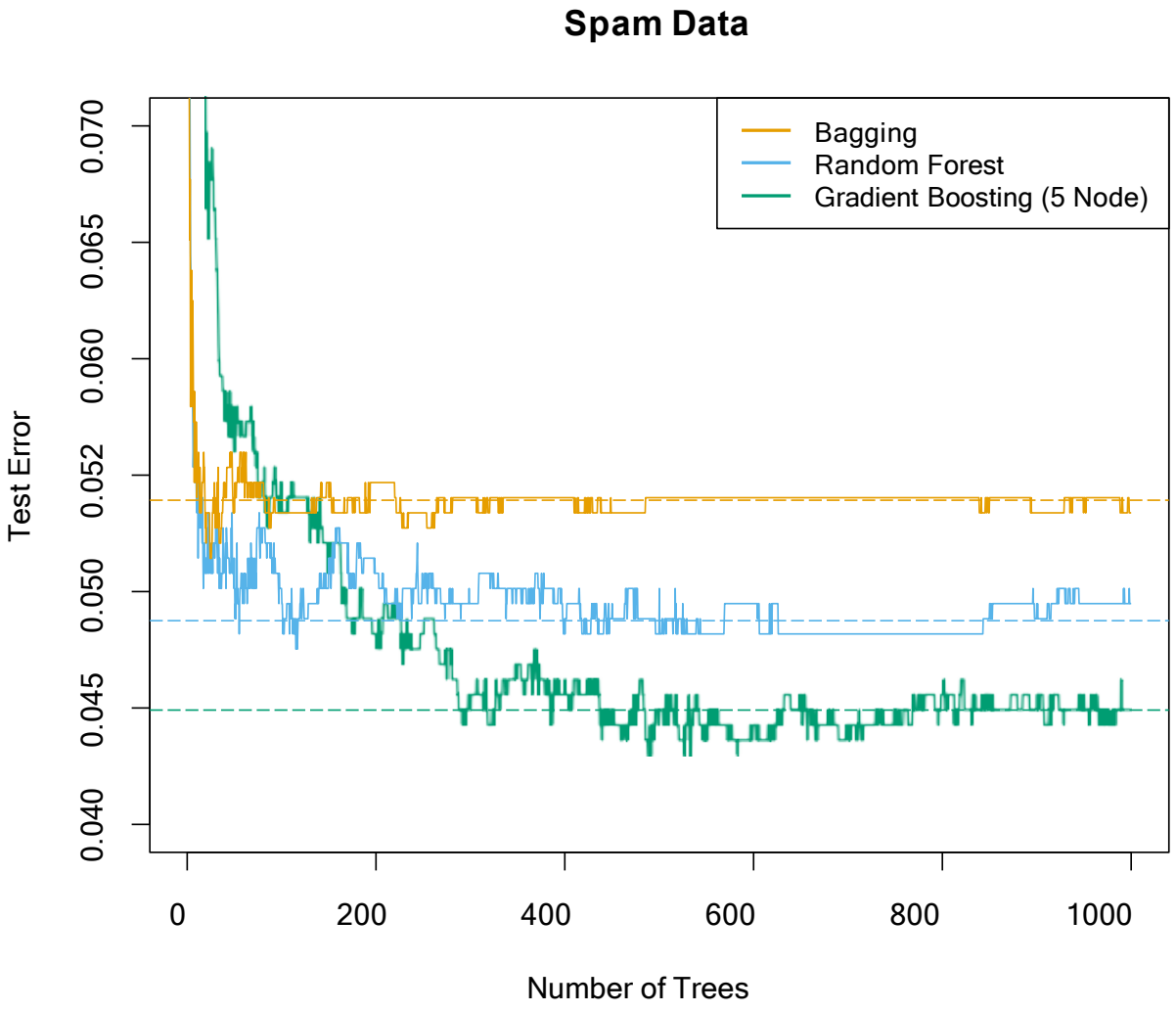
1. 木の数B。バギングやランダムフォレストとは異なり、Bが大きすぎると過学習が発生することがあるが、過学習が発生する場合でも、それは徐々に発生する傾向がある。Bの選択には交差検証を使用する。
2. 縮小パラメータλ、小さな正の数。これは、ブースティングが学習する速度を制御する。典型的な値は0.01または0.001であり、適切な選択は問題によって異なる。λが非常に小さい場合、良いパフォーマンスを達成するには非常に大きなBの値を使用する必要がある場合がある。
3. 各木の分割数dは、ブースティングの全体的な複雑さを制御する。たいていの場合d＝1がうまく機能する。その場合、各々の木が1つの分割のみからなる切り株である。このとき、各項が1つの変数のみを含むことから、ブースティング全体としては加法モデルの当てはめとなる。d分割には最大でd変数が含まれることがあるので、一般的には、dは交互作用の深さであり、ブースティングで得られたモデルの交互作用の次数を制御する。

回帰のもう一つの例



出典: *Elements of Statistical Learning, chapter 15.*

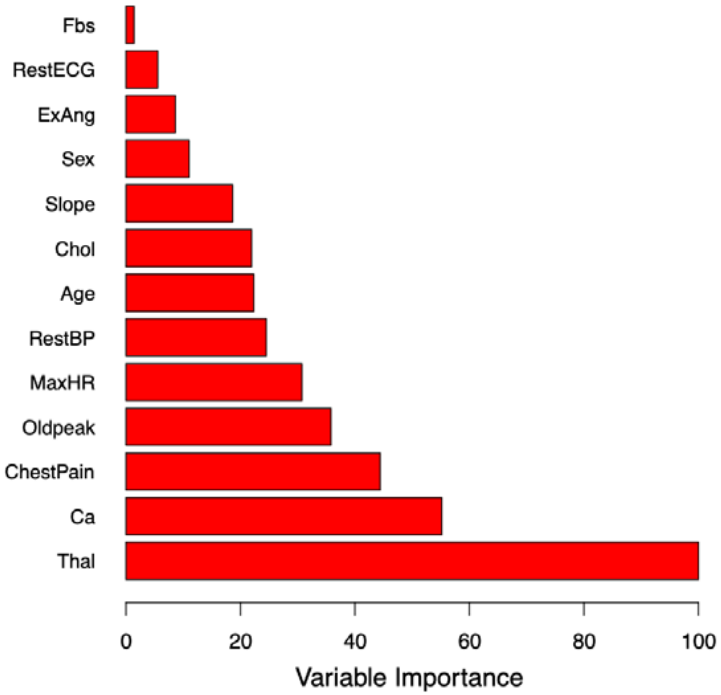
分類のもう一つの例



出典: *Elements of Statistical Learning, chapter 15.*

変数重要度の測定

- 回帰木のバギングやランダムフォレストの場合は、与えられた予測変数を分割することによって、RSSがどれだけ減少するかを計算し、B個の木について平均する。この値が大きければその変数は重要である。
- 同様に、分類木のバギングやランダムフォレストにおいても、与えられた変数を分割することによって、ジニ指数がどれだけ減少するかを計算し、B個の木について平均する。



心臓病データを用いて変数の重要度をプロットした図

まとめ

- 決定木は、回帰や分類のためのシンプルで解釈しやすいモデルである。
- しかし、予測精度の面では他の手法に及ばないことが多い。
- バギング、ランダムフォレスト、ブースティングは、木の予測精度を向上させるのに適した方法である。これらの手法は、学習データに対して多数の木を成長させ、得られた木のアンサンブルの予測値を組み合わせることで機能する。
- 後者のランダムフォレストとブースティングは、教師あり学習における最新の手法である。しかし、その結果の解釈は難しい。

第9章 サポートベクターマシン
-Support Vector Machines-

- 超平面と分離超平面
- 最大マージン分類器
- サポートベクター分類器
- 非線形性とカーネル
- カーネルとサポートベクターマシン
- サポートベクターマシンとロジスティック回帰

第9章 サポートベクターマシン
-Support Vector Machines-

2値分類問題を次のように考える。

データを2つのクラスに分類する平面を求める。

このような平面が存在しない場合には
- 分離の概念を緩める
- 特徴空間を拡張する

ことを考える。

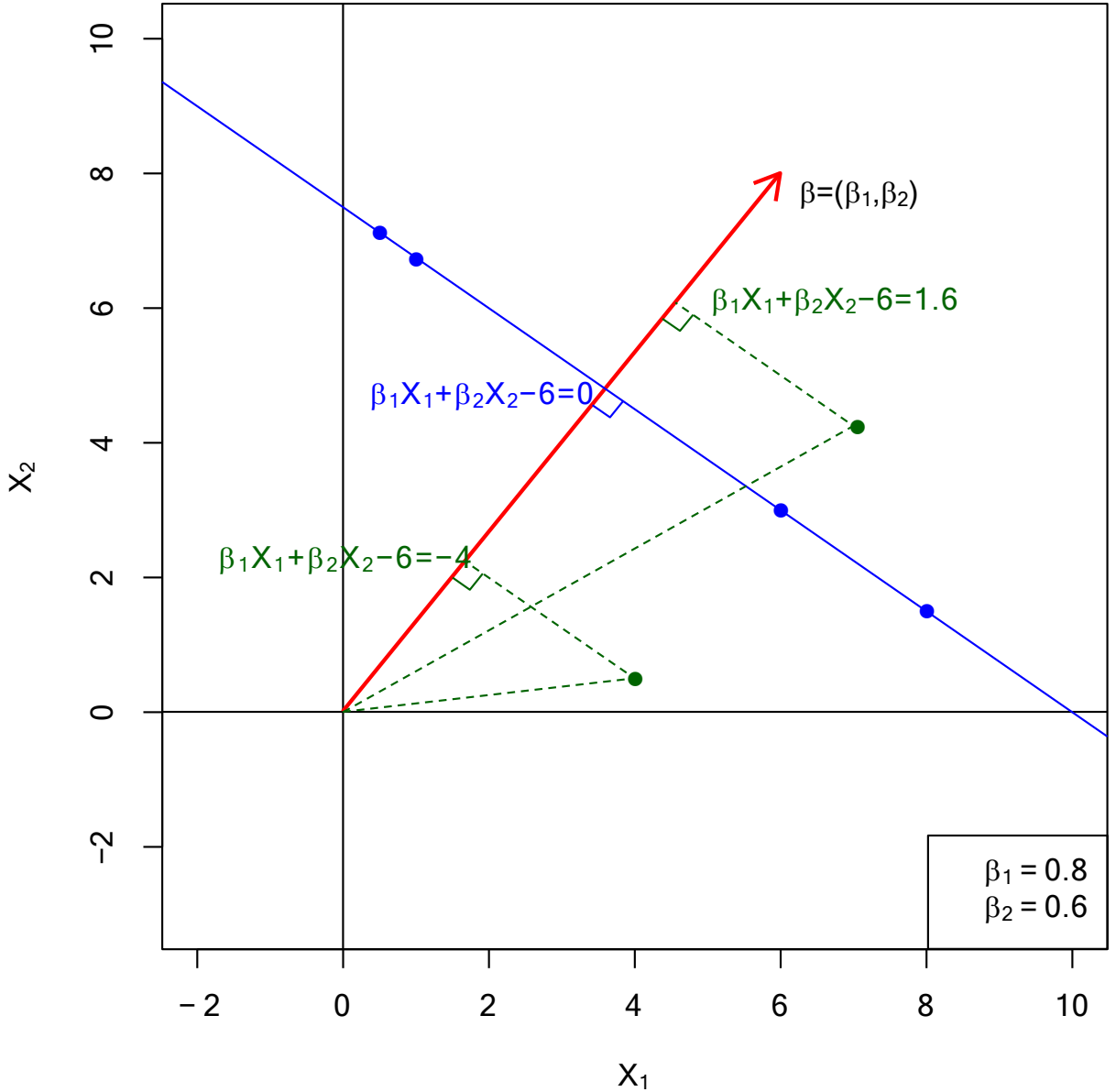
超平面とは何か

- p 次元空間における超平面は、$(p - 1)$ 次元のフラットなアフィン部分空間である。
- 一般に、超平面は次の式で定義できる

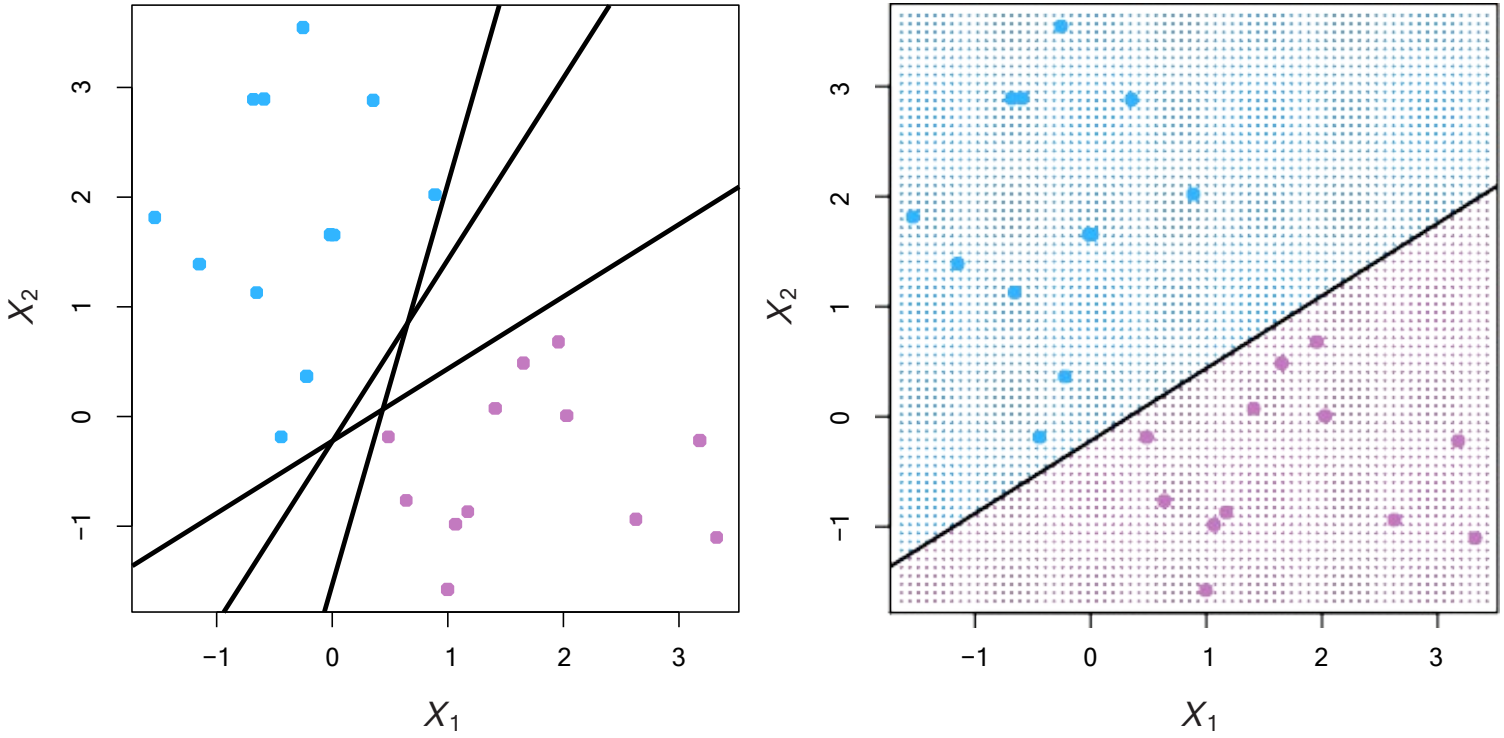
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p = 0$$

- $p = 2$の場合は超平面が直線である。
- $\beta_0 = 0$の場合は超平面が原点を通る。
- $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$は超平面の方向を決定する法線ベクトルである。

2次元における超平面


β=(β1,β2)
β1X1+β2X2−6=1.6
β1X1+β2X2−6=0
β1X1+β2X2−6=−4
β1 = 0.8
β2 = 0.6
X2
X1
10
8
6
4
2
0
−2
−2
0
2
4
6
8
10

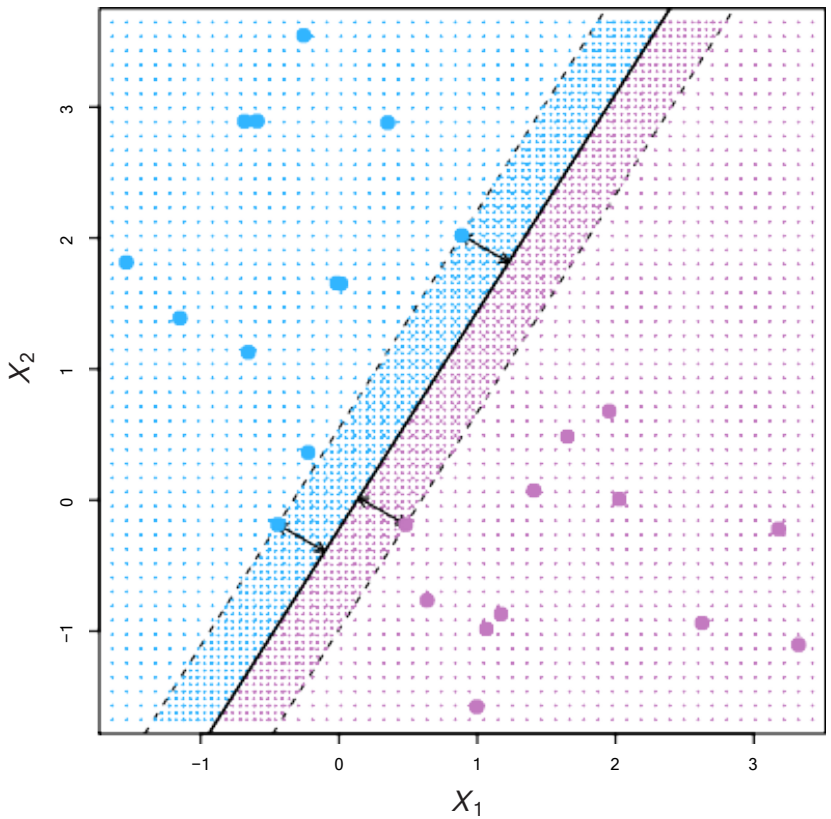
分離超平面



- $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$とする。もし$f(X) > 0$ならば$X$は超平面の片側に存在し、$f(X) < 0$ならば$X$は超平面をはさんで逆側に存在する。
- 青色のデータを$Y_i = +1$、紫色のデータを$Y_i = -1$とする。このとき、分離超平面$f(X) = 0$がすべてのiに対して$Y_i \cdot f(X_i) > 0$を満たす。

最大マージン分類器

無数に取りうる分離超平面のうち、2つのグループの間のギャップ（マージン）が最大となるような超平面を見出す。



以下の最適化問題を考える。

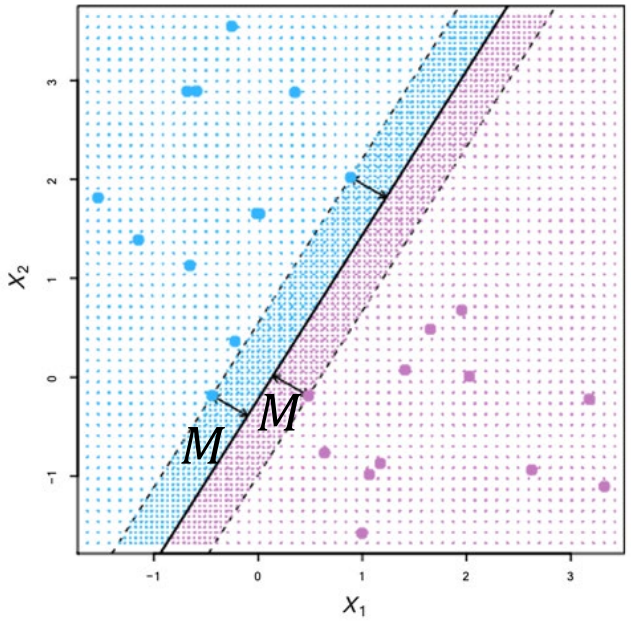
$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M$$

$$\text{for all } \; i = 1, \ldots, N.$$

訳者による補足1:



任意の点$(x_{i1}, \ldots, x_{ip})$から超平面$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = 0$までの距離は次のように表せる。

$$r(x_{i1}, \ldots, x_{ip}) = \frac{|\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}|}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}}$$

いま、i番目のデータに関して

$$y_i = \begin{cases} +1, i\text{番目のデータが青のクラスに属する} \\ -1, i\text{番目のデータが紫のクラスに属する} \end{cases}$$

とすると、分離超平面 $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = 0$ は、

$$\begin{cases} y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) > 0, y_i = +1\text{の場合} \\ y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) < 0, y_i = -1\text{の場合} \end{cases}$$

を満たしている。ここで、各観測点と分離超平面との距離の最小値を$M := \min_{i=1,..,N} r(x_{i1}, \ldots, x_{ip})$とすると、

$$M := \min_{i=1,..,N} \frac{|\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}|}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}}$$

と書ける（マージン）。

訳者による補足1–続き:

このとき、分離超平面は、すべてのiに対して

$$\begin{cases}\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\geq M, y_i=+1\text{の場合}\\ \beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\leq M, y_i=-1\text{の場合}\end{cases}$$

が成り立つものとする。この式は、

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq M, \forall i$$

と同値する。

実際に分離超平面が無数に存在するので、ここでは分離可能なデータに対して、マージンMが最大となる様、分離超平面を同定する。

$$\max_{\beta_0,\beta_1,\ldots,\beta_p}\left\{\min_{i=1,\ldots,N}\left\{\frac{\left|\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right|}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}\middle| i=1,..,N\right\}\middle| y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq M, i=1,\ldots,N\right\}$$

目的変数の絶対値をはずすと、

$$\max_{\beta_0,\beta_1,\ldots,\beta_p}\left\{\min_{i=1,\ldots,N}\left\{\frac{y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}\middle| i=1,..,N\right\}\middle| y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq M, i=1,\ldots,N\right\}$$

となり、さらに正規化すると、

$$\max_{\beta_0,\beta_1,\ldots,\beta_p}\left\{\min_{i=1,\ldots,N}\left\{y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\middle| \frac{i=1,\ldots,N,}{\sqrt{\beta_1^2+\cdots+\beta_p^2}=1}\right\}\middle| y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq M, i=1,\ldots,N\right\}$$

となる。ここで、制約条件$\sqrt{\beta_1^2+\cdots+\beta_p^2}=1$のもとで

$\min_{i=1,\ldots,N}\left\{y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\middle| i=1,..,N\right\}=M$ となることに注意すると、

訳者による補足1–続き:

マージンMの最大化問題は

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \left\{ M \;\middle|\; \begin{array}{c} y_i(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}) \geq M, i=1,\ldots,N \\ \sqrt{\beta_1^2+\cdots+\beta_p^2}=1 \end{array} \right\}$$

として定式化できる。

$\sqrt{\beta_1^2+\cdots+\beta_p^2}=1$は$\sum_{j=1}^{p}\beta_j^2=1$と等価だから、上記の問題は最終的に次のように表せる。

$$\begin{aligned} &\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M \\ &\text{subject to } \sum_{j=1}^{p}\beta_j^2=1, \\ &y_i(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}) \geq M \\ &\text{for all } \; i=1,\ldots,N. \end{aligned}$$

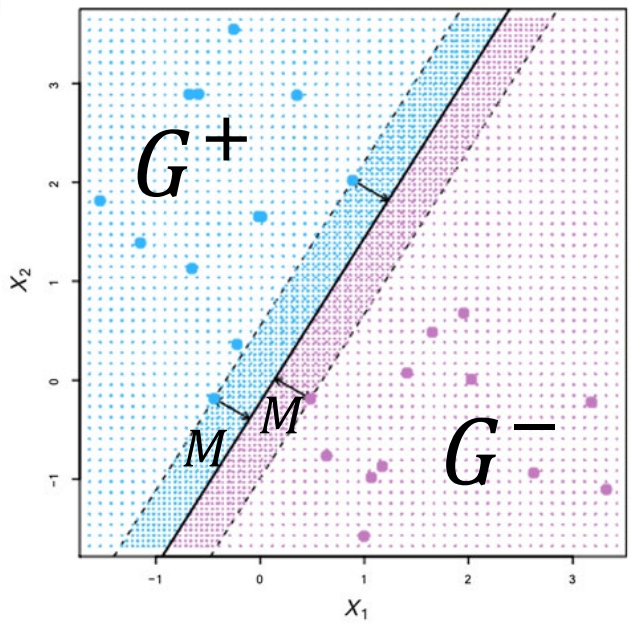
訳者による補足2–別の考え方:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M$$

$$\text{for all} \quad i = 1, \ldots, N.$$



スタート:分離超平面までの最小距離を最大化する問題を考える。

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \min_{i=1,..,N} \frac{\left|\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right|}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}}$$

$$s.t.$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0, i \in G^+ : group\ of\ blue\ points$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ points$$

訳者による補足2–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \min_{i=1,..,N} \frac{\left|\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right|}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}}$$

$$s.t.$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0, i \in G^+ : group\ of\ blue\ points$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$$

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \min_{i} \left\{ \left(\frac{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}} \right)_{i \in G^+}, -\left(\frac{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}} \right)_{i \in G^-} \right\}$$

$$s.t.$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0, i \in G^+ : group\ of\ blue\ points$$

$$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$$

訳者による補足2–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \min_i \left\{ \left(\frac{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}} \right)_{i \in G^+}, -\left(\frac{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}} \right)_{i \in G^-} \right\}$$

$s.t.$

$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0, i \in G^+ : group\ of\ blue\ points$

$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{\sqrt{\beta_1^2 + \cdots + \beta_p^2}} \min_i \left\{ \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)_{i \in G^+}, -\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right)_{i \in G^-} \right\}$$

$s.t.$

$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} > 0, i \in G^+ : group\ of\ blue\ points$

$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$

訳者による補足2–続き：

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{\sqrt{\beta_1^2+\cdots+\beta_p^2}} \min_i \left\{ \left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)_{i\in G^+}, -\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)_{i\in G^-} \right\}$$

$$s.t.$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} > 0, i \in G^+ : group\ of\ blue\ points$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$$

$\xi^+ := \min\{\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} | i \in G^+\}, -\xi^- := \min\{-(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}) | i \in G^-\}$とすると、

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{\sqrt{\beta_1^2+\cdots+\beta_p^2}} \min\{\xi^+, -\xi^-\}$$

$$s.t.$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \geq \xi^+ > 0, i \in G^+ : group\ of\ blue\ points$$

$$\xi^- \leq \beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$$

訳者による補足2–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{\sqrt{\beta_1^2+\cdots+\beta_p^2}} \min\{\xi^+, -\xi^-\}$$

$$s.t.$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \geq \xi^+ > 0, i \in G^+ : group\ of\ blue\ points$$

$$\xi^- \leq \beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} < 0, i \in G^- : group\ of\ mauve\ point$$

ここで、$M := \xi^+ = -\xi^- > 0$とすると、上記の問題は次の問題と同値である。

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \geq M, i \in G^+ : group\ of\ blue\ points$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \leq M, i \in G^- : group\ of\ mauve\ point$$

$$M > 0$$

訳者による補足2–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \geq M, i \in G^+ : group\ of\ blue\ points$$

$$\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip} \leq M, i \in G^- : group\ of\ mauve\ point$$

$$M > 0$$

さらに、$y_i = \begin{cases} +1, i\text{番目のデータが青のクラスに属する} \\ -1, i\text{番目のデータが紫のクラスに属する} \end{cases}$

とすると、

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) \geq M, \forall\, i = 1, \ldots, N$$

$$M > 0$$

訳者による補足2–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) \geq M, \forall\, i=1,\ldots,N$$

$$M>0$$

ここで正規化条件として$\sqrt{\beta_1^2+\cdots+\beta_p^2}=1$とおくと、

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

$$\text{subject to } \sum_{j=1}^{p}\beta_j^2=1,$$

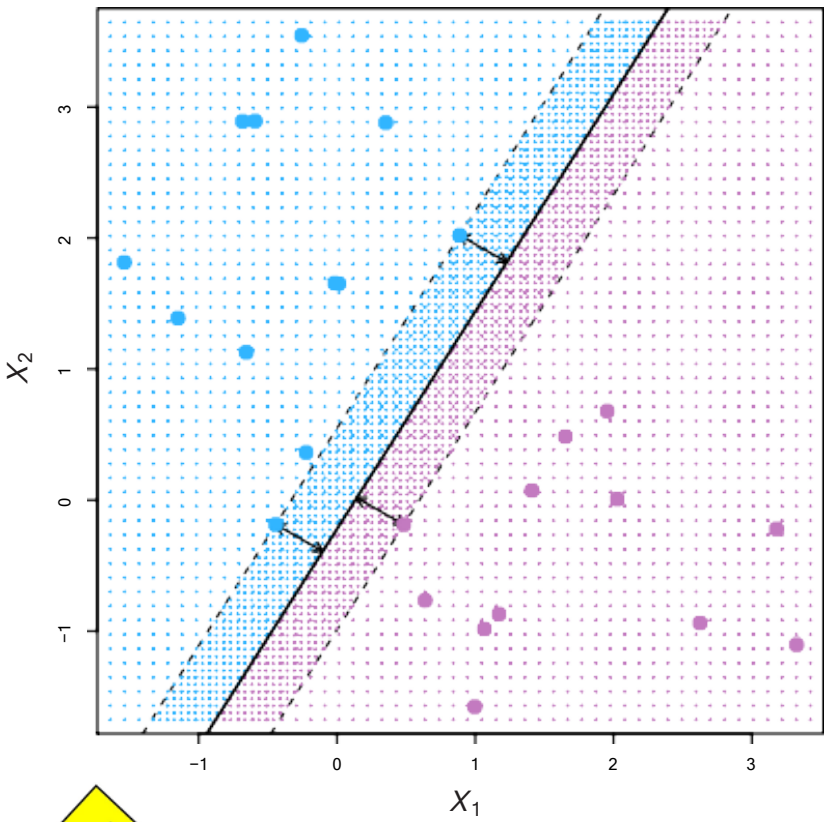
$$y_i(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}) \geq M$$

$$\text{for all } \; i=1,\ldots,N.$$

$$M>0$$

最大マージン分類器

無数に取りうる分離超平面のうち、2つのグループの間のギャップ（マージン）が最大となるような超平面を見出す。



以下の最適化問題を考える。

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

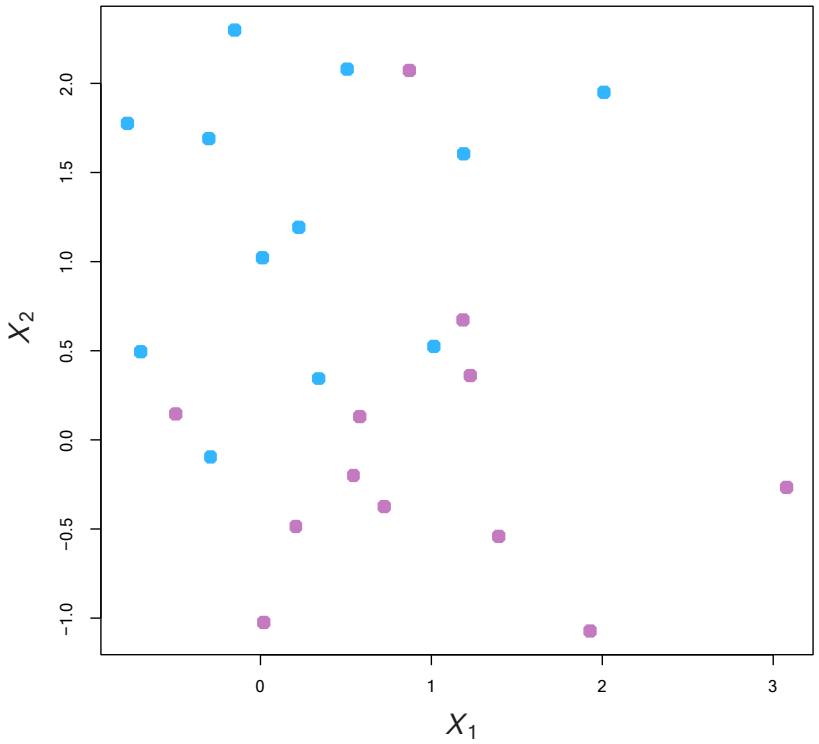
$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \geq M$$

$$\text{for all} \quad i = 1, \ldots, N.$$

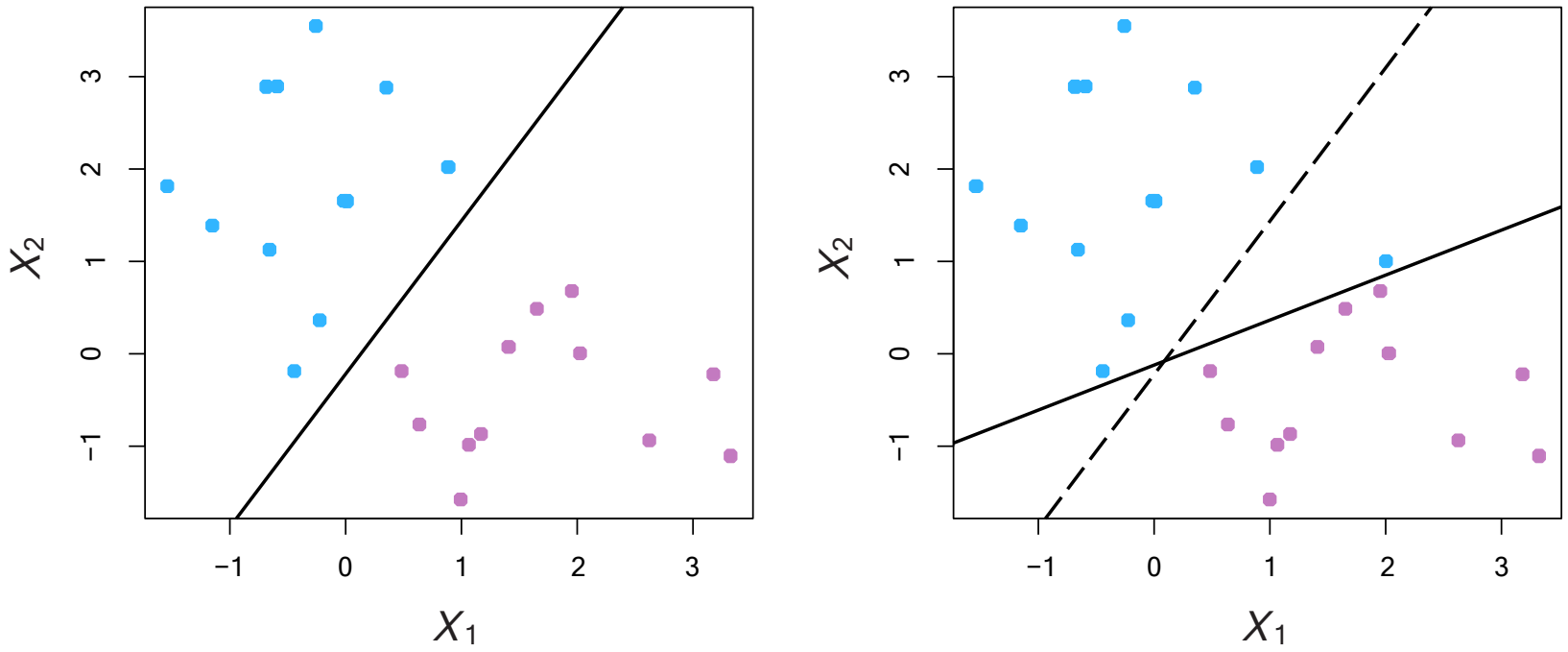
この問題は凸2次計画問題として定式化できる。Rパッケージe1071のsvm()関数を用いて効率的に解くことができる

分離不可能なデータ



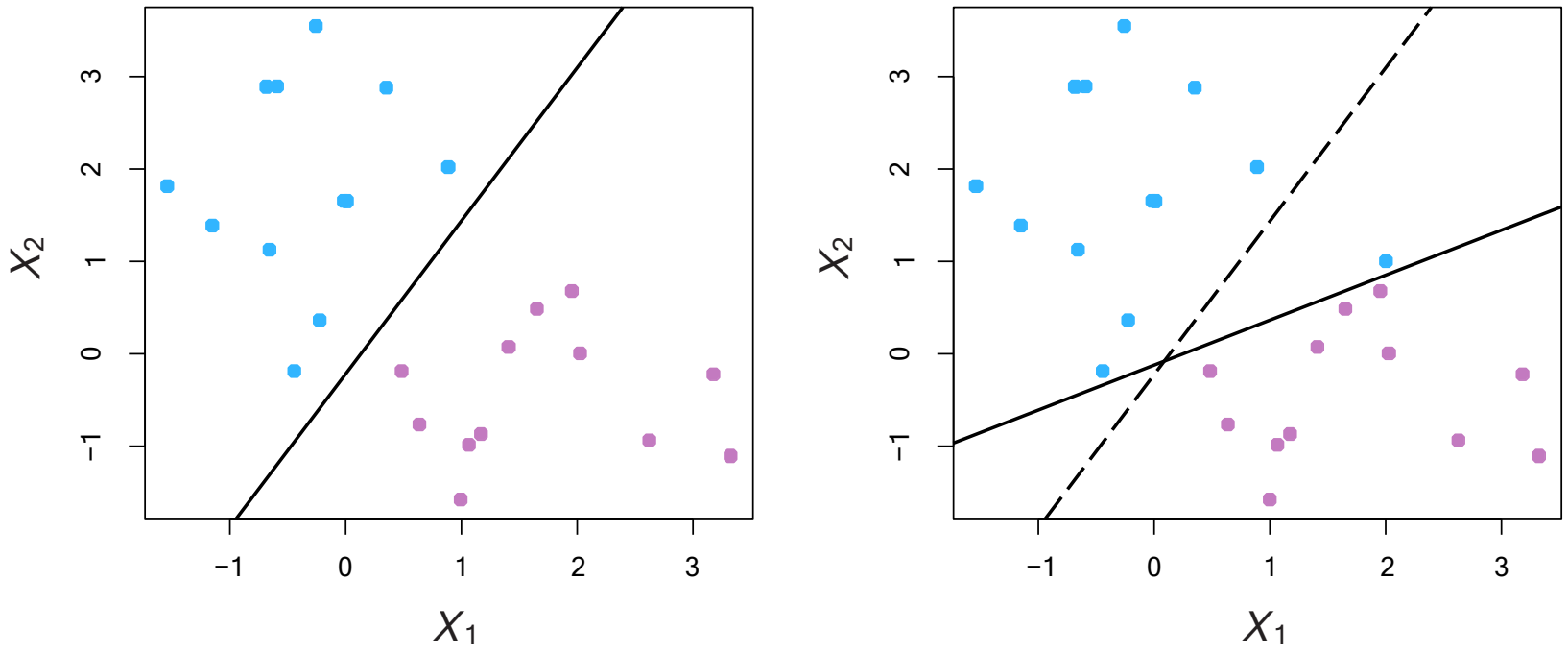
左図にあるようなデータは線形直線で完全に分離するとこができない。
これは$N < p$が満たさなければ稀なケースではない。

ノイズのあるデータ



たとえ分離超平面が存在しているとしても、分離超平面に基づき分類が適切でない場合もある。

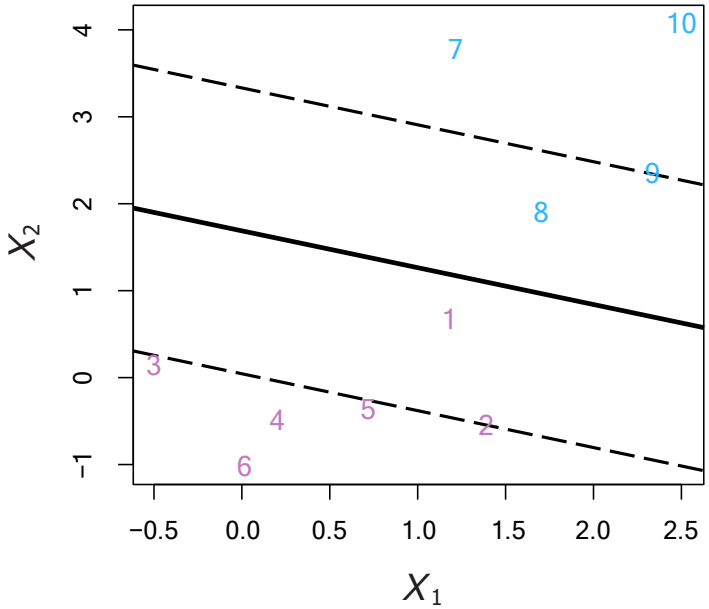
ノイズのあるデータ

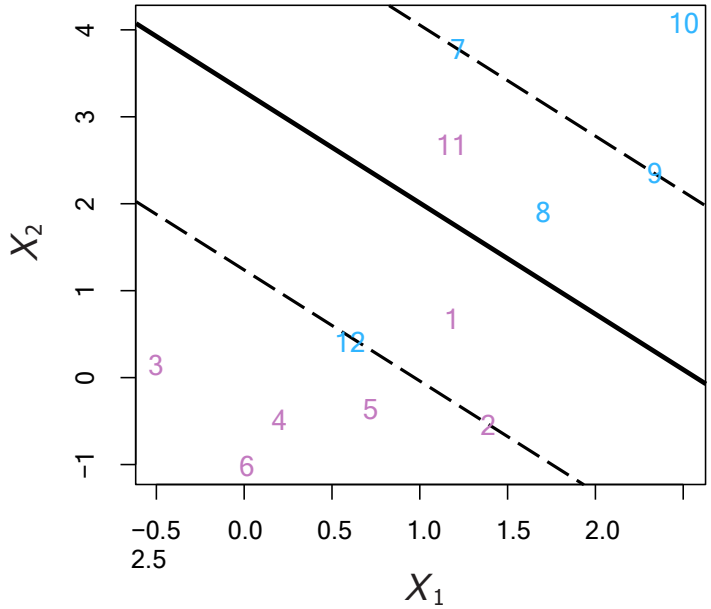


たとえ分離超平面が存在しているとしても、分離超平面に基づき分類が適切でない場合もある。

サポートベクター分類器はソフトマージンを最大化している。

サポートベクター分類器





$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

訳者による補足3:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) \geq M, \forall\, i=1,\ldots,N$$

$$M>0$$

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M/M}{\sqrt{\left(\frac{\beta_1}{M}\right)^2+\cdots+\left(\frac{\beta_p}{M}\right)^2}}$$

$$s.t.$$

$$y_i\left(\frac{\beta_0}{M}+\frac{\beta_1}{M}x_{i1}+\cdots+\frac{\beta_p}{M}x_{ip}\right) \geq \frac{M}{M}, \forall\, i=1,\ldots,N$$

$$\frac{M}{M}>0$$

訳者による補足3–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{M/M}{\sqrt{\left(\frac{\beta_1}{M}\right)^2+\cdots+\left(\frac{\beta_p}{M}\right)^2}}$$

$$s.t.$$

$$y_i\left(\frac{\beta_0}{M}+\frac{\beta_1}{M}x_{i1}+\cdots+\frac{\beta_p}{M}x_{ip}\right)\geq\frac{M}{M},\forall\, i=1,\ldots,N$$

$$\frac{M}{M}>0$$

ここで、$\beta_0 \coloneqq \frac{\beta_0}{M}, \beta_1 \coloneqq \frac{\beta_1}{M}, \ldots, \beta_p \coloneqq \frac{\beta_p}{M}$とおくと、上記の問題は次の問題に変換される。

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}\right)\geq 1,\forall\, i=1,\ldots,N$$

訳者による補足3–続き:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq 1, \forall\, i=1,\ldots,N$$

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} \frac{2}{\sqrt{\beta_1^2+\cdots+\beta_p^2}}$$

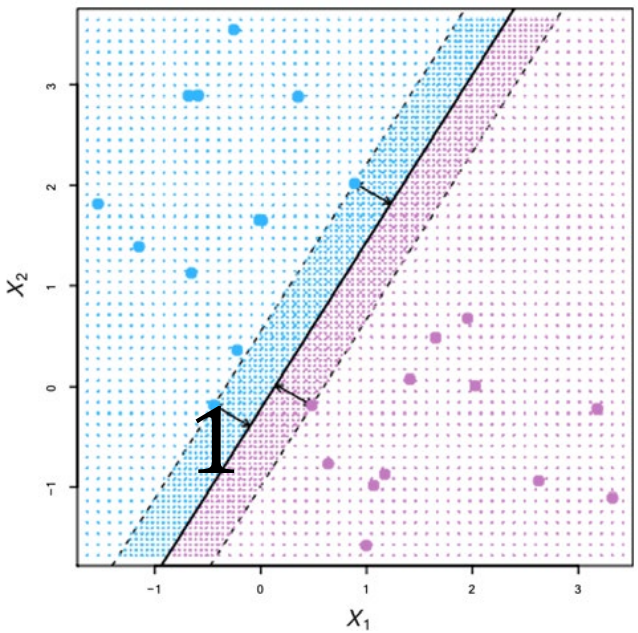
$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq 1, \forall\, i=1,\ldots,N$$

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq 1, \forall\, i=1,\ldots,N$$



訳者による補足3–続き：

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}$$
$$s.t.$$
$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq 1, \forall\, i=1,\ldots,N$$

i番目のデータの（マージンを）はみ出した長さを$\epsilon_i \geq 0$とし、すべてのiについてはみ出した長さの合計が$C \geq 0$以下であると仮定すると、

$$\min_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_N} \frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}$$
$$s.t.$$
$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq 1-\epsilon_i, \forall\, i=1,\ldots,N$$
$$\sum_{i=1}^{N}\epsilon_i \leq C$$
$$\epsilon_i \geq 0, \forall\, i$$

訳者による補足3–続き:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_N} \frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) \geq 1-\epsilon_i, \forall\, i=1,\ldots,N$$

$$\sum_{i=1}^{N} \epsilon_i \leq C$$

$$\epsilon_i \geq 0, \forall\, i$$

上記の問題は、パラメータCと$\gamma > 0$を適切に対応させれば、次の問題と等価になる。

$$\min_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_N} \frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}+\gamma\sum_{i=1}^{N}\epsilon_i$$

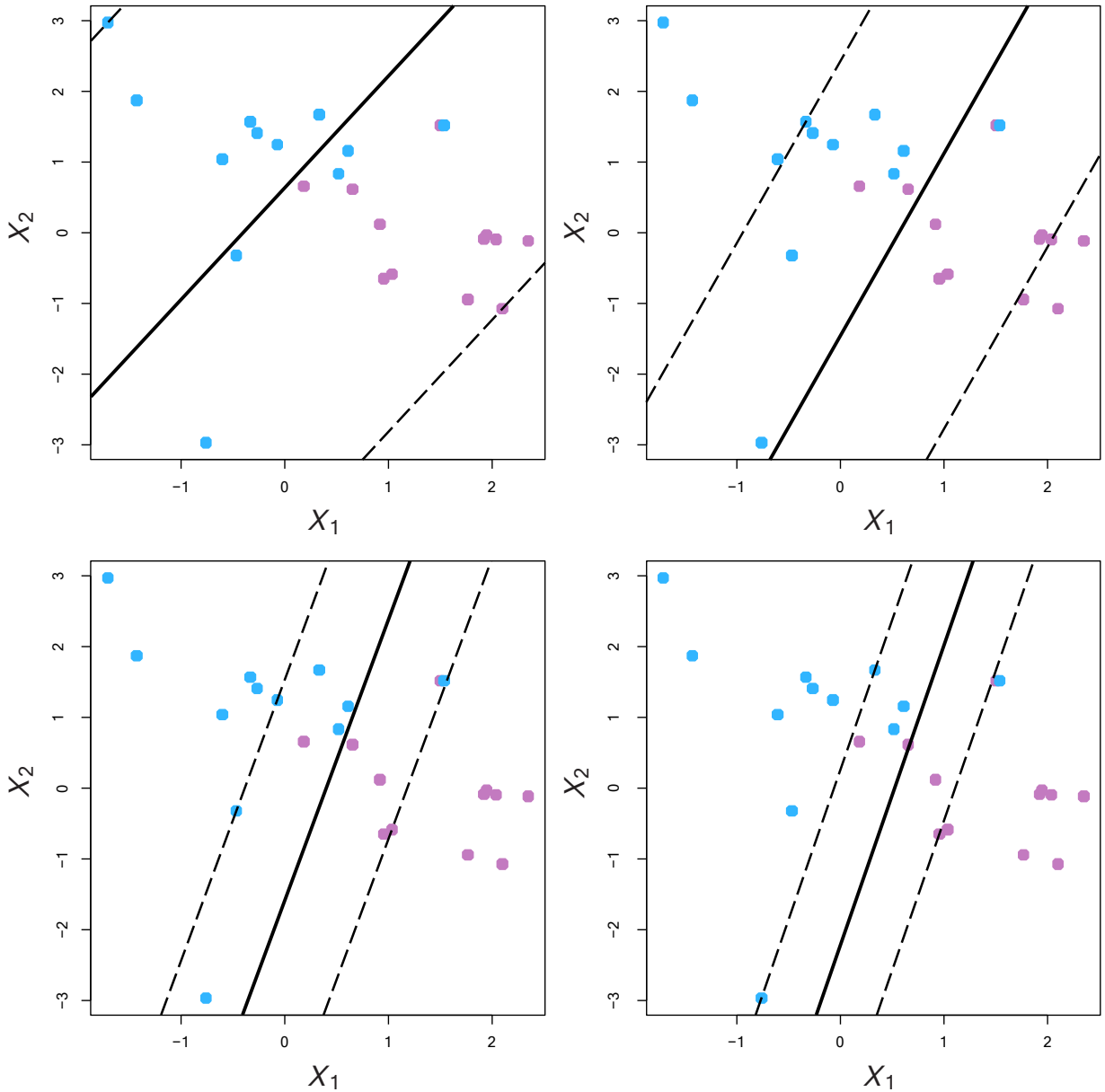
$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right) \geq 1-\epsilon_i, \forall\, i=1,\ldots,N$$

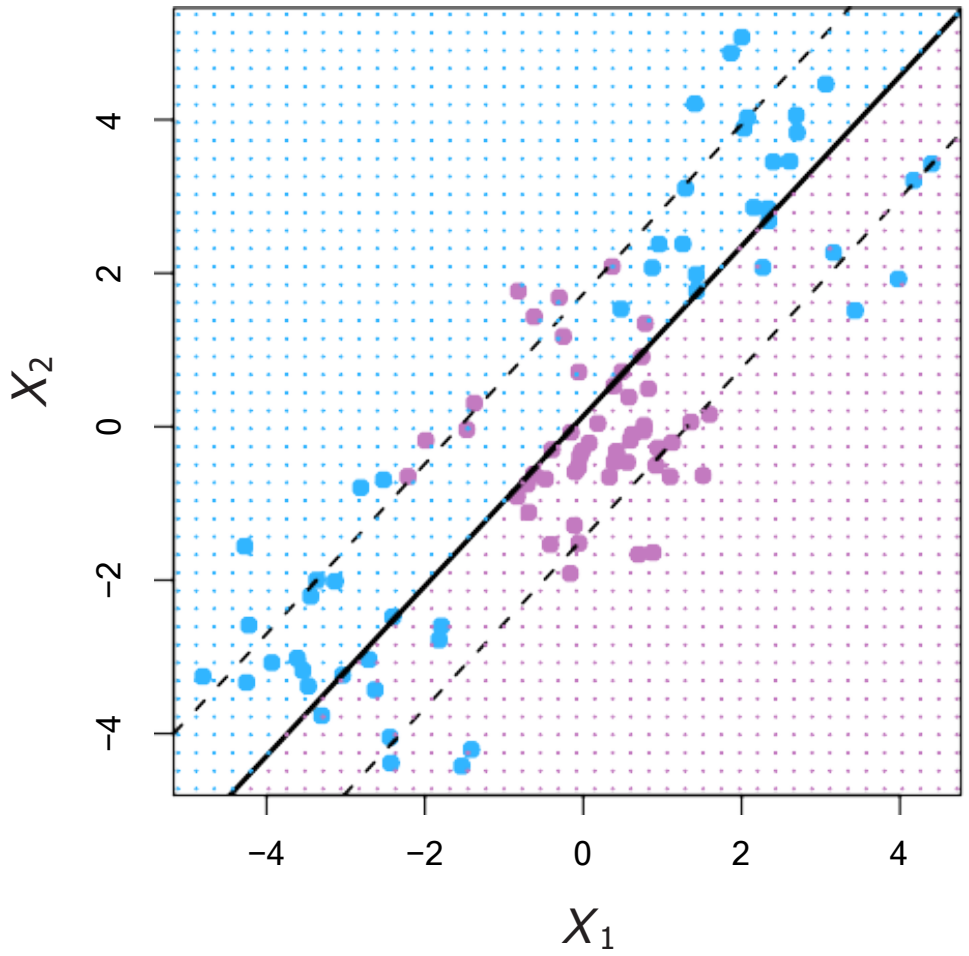
$$\epsilon_i \geq 0, \forall\, i$$

この問題は、2次計画問題である。

正則化パラメータC



線形の境界による誤分類



どのようなCを用いても線形の境界でうまく分類できないことがある。

特徴空間の拡張

- $X_1^2, X_1^3, X_1X_2, X_1X_2^2$ …のような変換を用いて特徴空間を大きくする。
- 拡張された特徴空間の中でサポートベクター分類器を適用する。
- これが非線形の決定境界を導く。

特徴空間の拡張

- $X_1^2, X_1^3, X_1X_2, X_1X_2^2$ …のような変換を用いて特徴空間を大きくする。
- 拡張された特徴空間の中でサポートベクター分類器を適用する。
- これが非線形の決定境界を導く。

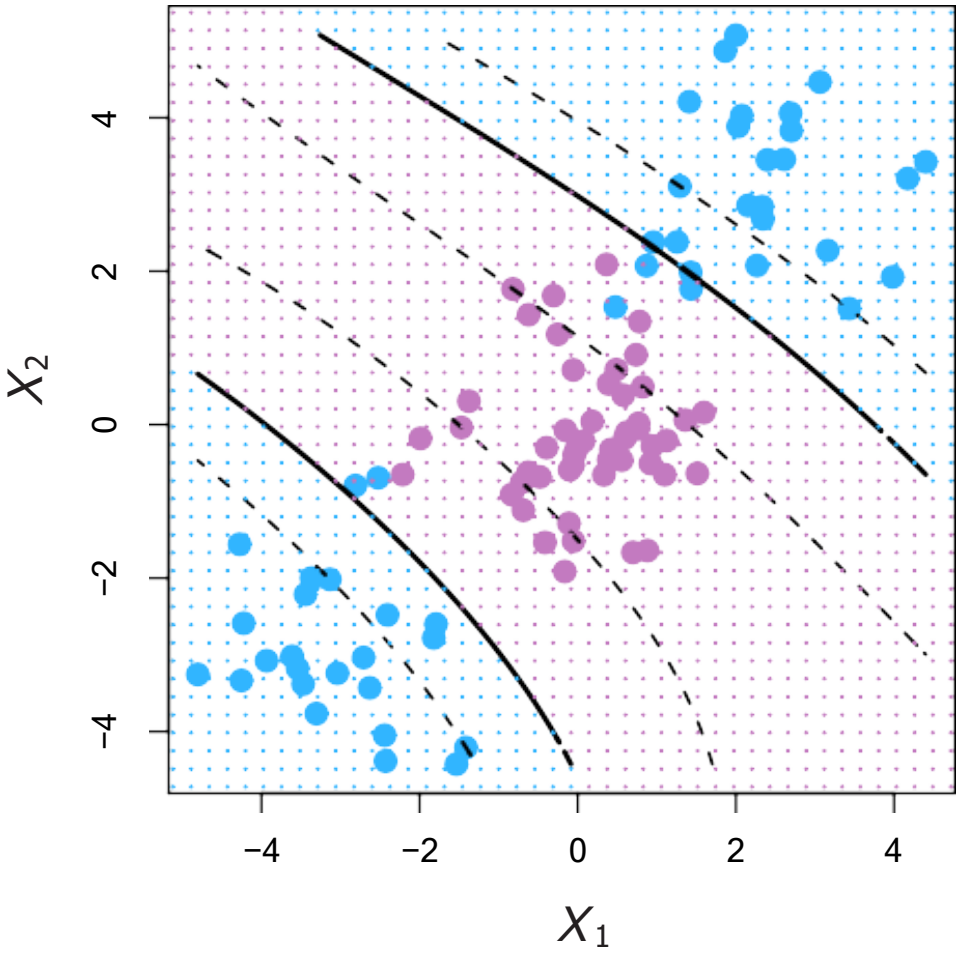
例：
(X_1, X_2)のかわりに$(X_1, X_2, X_1^2, X_2^2, X_1X_2)$を用いてサポートベクター分類器に適用する。このとき、決定境界は

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 = 0$$

となる。
元の特徴空間において、決定境界は非線形である。

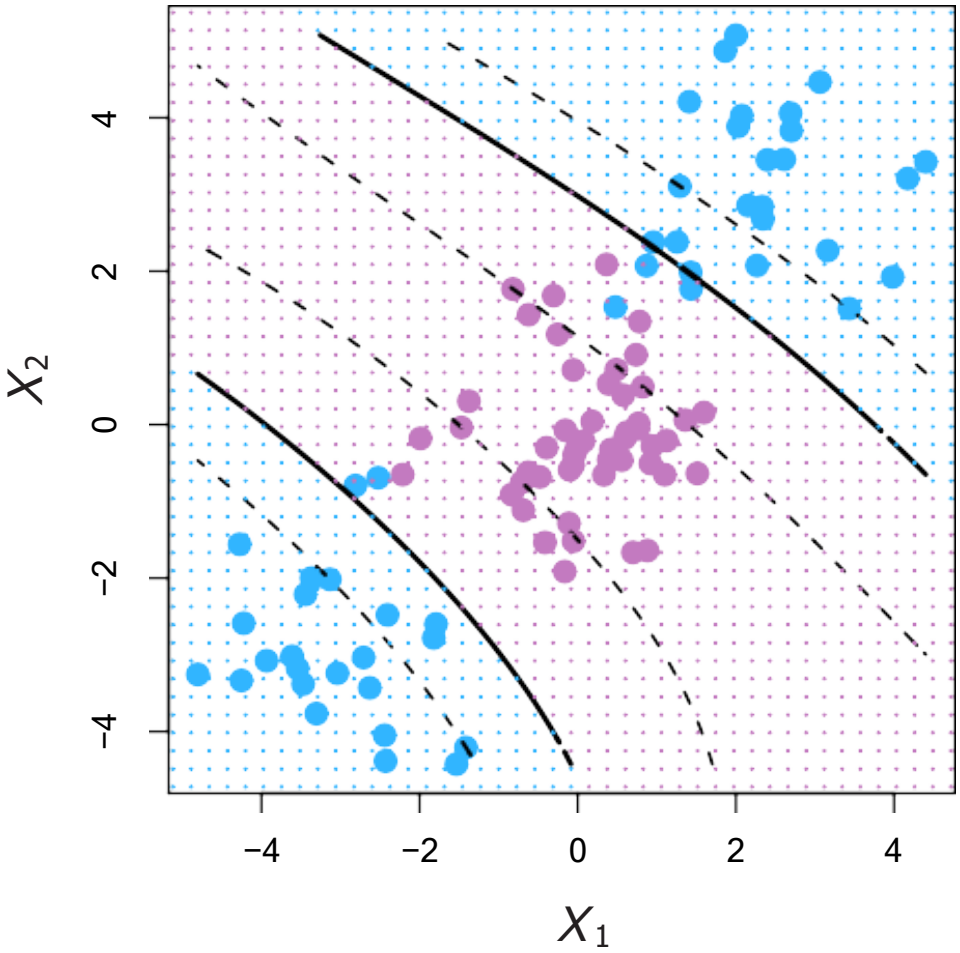
3次多項式

3次多項式によって張られた高次元空間の例を示す。
このとき、変数が2個から9個になる。
拡張された特徴空間におけるサポートベクター分類器が低次元空間の問題を解いている。



3次多項式

3次多項式によって張られた高次元空間の例を示す。
このとき、変数が2個から9個になる。
拡張された特徴空間におけるサポートベクター分類器が低次元空間の問題を解いている。



$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \beta_6 X_1^3 + \beta_7 X_2^3 + \beta_8 X_1 X_2^2 + \beta_9 X_1^2 X_2 = 0$$

非線形性とカーネル

- 多項式（特に高次元のもの）は性能が悪い。
- サポートベクター分類器に非線形性を導入するより優雅で制御された方法がある。それはカーネルを使うことである。
- これらについて説明する前に、サポートベクター分類器における内積の役割を理解する必要がある。

内積とサポートベクター

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

内積

内積とサポートベクター

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$ 内積

線形のサポートベクター分類器は次のように表せる。

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x_i, x_{i'} \rangle$$ n個のパラメータ

訳者による補足4:

スタート:ソフトマージン最大化問題の双対問題を考える。
以下の問題を主問題とする。

$$\min_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_N} \frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}+\gamma\sum_{i=1}^{N}\epsilon_i$$

$$s.t.$$

$$y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)\geq 1-\epsilon_i, \forall\, i=1,\ldots,N$$

$$\epsilon_i \geq 0, \forall\, i$$

ここで、非負の係数$\alpha_1,\ldots,\alpha_N,\mu_1,\ldots,\mu_N$を導入すると、以下のラグランジュ緩和問題が得られる。

$$\min L\left(\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_N;\alpha_1,\ldots,\alpha_N,\mu_1,\ldots,\mu_N\right)$$

$$:=\frac{1}{2}\sqrt{\beta_1^2+\cdots+\beta_p^2}+\gamma\sum_{i=1}^{N}\epsilon_i-\sum_{i=1}^{N}\alpha_i\left\{y_i\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}\right)-(1-\epsilon_i)\right\}-\sum_{i=1}^{N}\mu_i\epsilon_i$$

訳者による補足4–続き:

$$\min L(\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_N; \alpha_1, \ldots, \alpha_N, \mu_1, \ldots, \mu_N)$$
$$:= \frac{1}{2}\sqrt{\beta_1^2 + \cdots + \beta_p^2} + \gamma \sum_{i=1}^{N} \epsilon_i - \sum_{i=1}^{N} \alpha_i \{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - (1 - \epsilon_i)\} - \sum_{i=1}^{N} \mu_i \epsilon_i$$

このラグランジュ緩和問題は制約なし最適化問題なので、その最適解は$L(\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_N; \alpha_1, \ldots, \alpha_N, \mu_1, \ldots, \mu_N)$の$\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_N$についての偏微分をゼロにすることで求められる。

$$0 = \sum_{i=1}^{N} \alpha_i y_i$$

$$\beta_m = \sum_{i=1}^{N} \alpha_i y_i x_{im}, m = 1, \ldots, p$$

$$\gamma - \alpha_i - \mu_i = 0, i = 1, \ldots, N$$

これらの条件を $L(\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_N; \alpha_1, \ldots, \alpha_N, \mu_1, \ldots, \mu_N)$ に代入すると、次の双対問題が得られる。

訳者による補足4–続き:

$$\max_{\alpha_1,\ldots,\alpha_N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

$$s.t.$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq \gamma, i = 1, \ldots, N$$

この双対問題の解を$\hat{\alpha}_1, \ldots, \hat{\alpha}_N$とすると、分離超平面の係数と切片の推定値はそれぞれ、

$$\hat{\beta}_m = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_{im}, m = 1, \ldots, p$$

$$\hat{\beta}_0 = -\frac{1}{2}\left(\sum_{m=1}^{p} \hat{\beta}_m (x_{im})_{i \in H^+} + \sum_{m=1}^{p} \hat{\beta}_m (x_{im})_{i \in H^-} \right)$$

となる。ここで、H^+, H^-は分離超平面をはさむ2つの超平面を表す。

訳者による補足4–続き:

この双対問題の解を$\hat{\alpha}_1, \ldots, \hat{\alpha}_N$とすると、分離超平面の係数と切片の推定値はそれぞれ、

$$\hat{\beta}_m = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_{im}\,, m = 1, \ldots, p$$

$$\hat{\beta}_0 = -\frac{1}{2}\left(\sum_{m=1}^{p} \hat{\beta}_m (x_{im})_{i \in H^+} + \sum_{m=1}^{p} \hat{\beta}_m (x_{im})_{i \in H^-} \right)$$

となる。ここで、H^+, H^-は分離超平面をはさむ2つの超平面を表す。
このとき、分離超平面$f(x) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = 0$の関数の部分は次のように表せる。

$$\hat{f}(x) = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_m x_{im}$$

$$= \hat{\beta}_0 + \sum_{i=1}^{m} \left(\sum_{i=1}^{N} \hat{\alpha}_i y_i x_{im} \right) x_{i'm}$$

$$= \hat{\beta}_0 + \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle x_i, x_{i'} \rangle$$

内積とサポートベクター

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$ 内積

線形のサポートベクター分類器は次のように表せる。

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x_i, x_{i'} \rangle$$ n個のパラメータ

パラメータβ_0と$\alpha_1, \ldots, \alpha_N$を推定するために必要なものは、訓練データのすべての組み合わせの$\binom{n}{2} = \frac{n(n-1)}{2}$個の内積$\langle x_i, x_{i'} \rangle$のみである。

内積とサポートベクター

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$ 内積

線形のサポートベクター分類器は次のように表せる。

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x_i, x_{i'} \rangle$$ n個のパラメータ

パラメータβ_0と$\alpha_1, \dots, \alpha_N$を推定するために必要なものは、訓練データのすべての組み合わせの$\binom{n}{2} = \frac{n(n-1)}{2}$個の内積$\langle x_i, x_{i'} \rangle$のみである。また、$\alpha_i$はサポートベクターにおいてのみ非ゼロであるため、ここでサポートベクターの添字の集合をSとすると、上記の式は次のように書き直すことができる。

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x_i, x_{i'} \rangle$$

カーネルとサポートベクターマシン

- 観測値間の内積を計算できれば、SV分類器を適合させることができる。
- 一部の特殊なカーネル関数を用いることが可能である。

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

- この関数は、次数dの多項式カーネルと呼ばれる。

カーネルとサポートベクターマシン

- 観測値間の内積を計算できれば、SV分類器を適合させることができる。
- 一部の特殊なカーネル関数を用いることが可能である。

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$

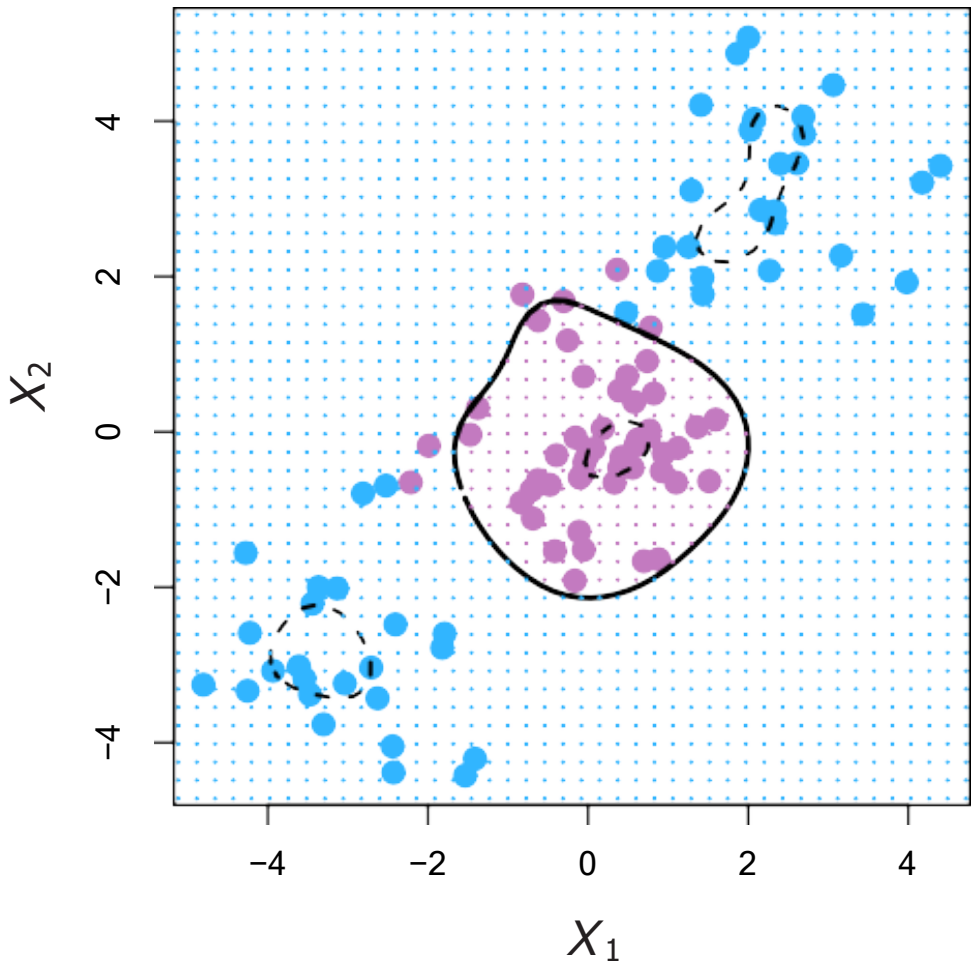
- この関数は、次数dの多項式カーネルと呼ばれる。

- この関数を利用した場合の解は以下に示す。

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i K(x, x_i).$$

ラジアルカーネル

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

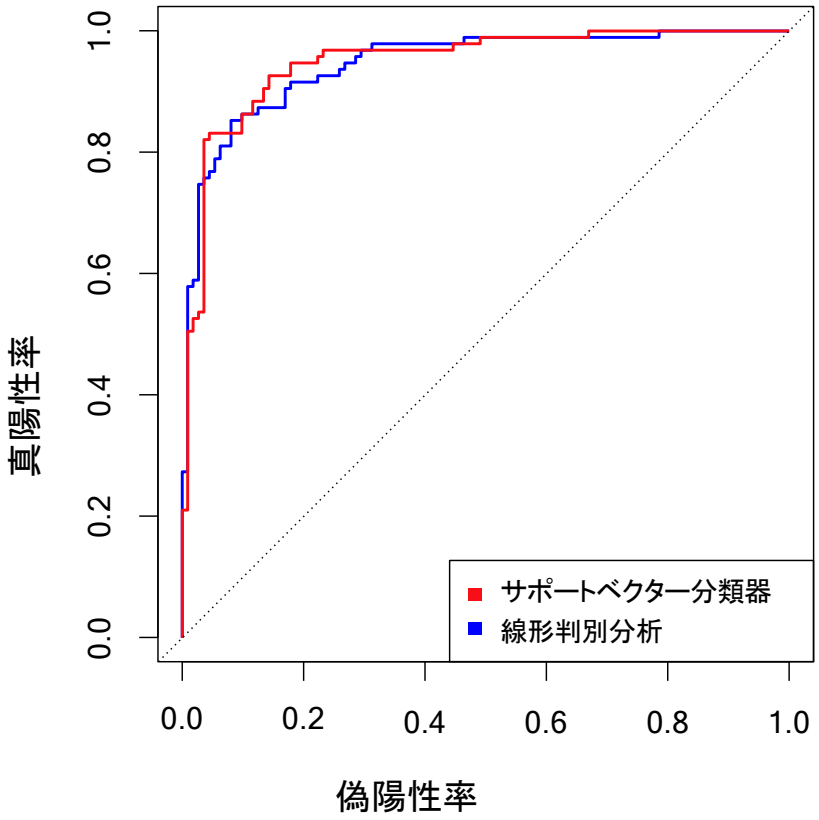


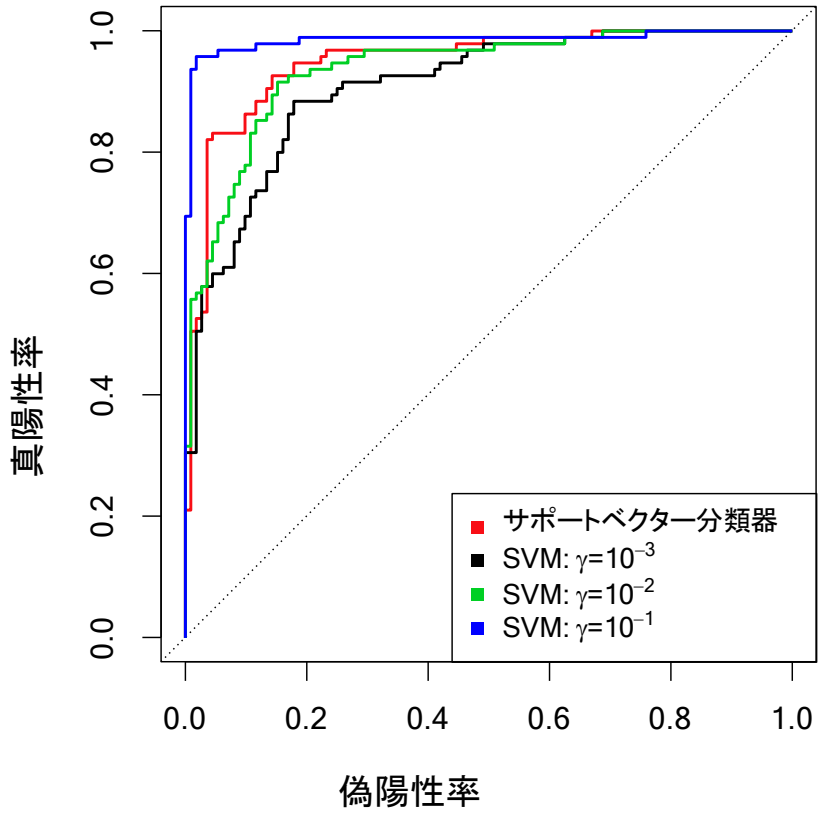
$$\hat{f}(x) = \hat{\beta}_0 + \sum_{i \in S} \hat{\alpha}_i \, K(x_i, x_{i'})$$

特徴写像を陽に扱うことなく、高次元空間上の計算を回避できる。

次元を縮小することで分散を抑制できる。

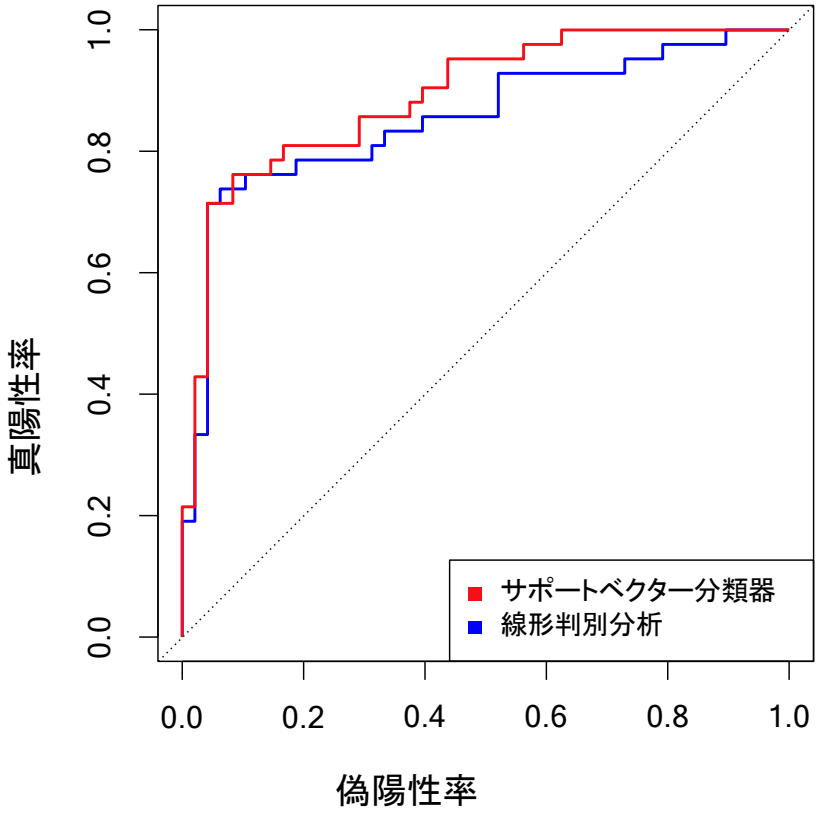
例：心臓病データ

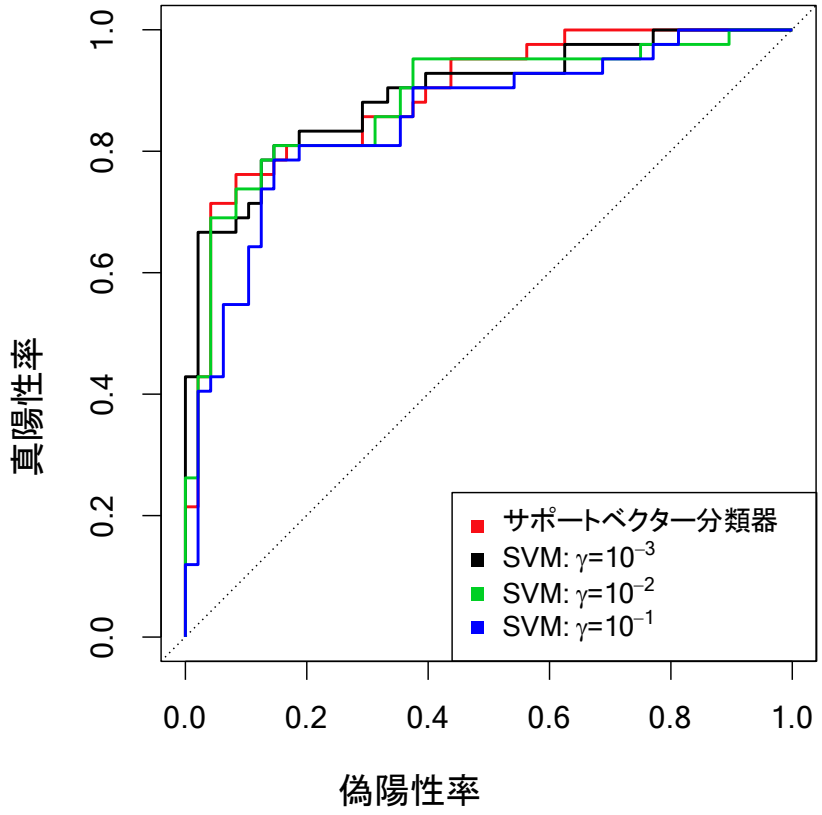




ROC曲線は様々な閾値 t に対する偽陽性率と真陽性率を示している。

例–続き：心臓病データ


真陽性率
偽陽性率
0.0
0.2
0.4
0.6
0.8
1.0
サポートベクター分類器
線形判別分析


真陽性率
偽陽性率
0.0
0.2
0.4
0.6
0.8
1.0
サポートベクター分類器
SVM: γ=10^−3
SVM: γ=10^−2
SVM: γ=10^−1

SVMs: 3つ以上のクラスがある場合は？

これまで、2つのクラスにおける分類の場合（$K = 2$）について議論してきた。$K > 2$の場合はどのように対処すべきか？

SVMs: 3つ以上のクラスがある場合は？

これまで、2つのクラスにおける分類の場合（$K = 2$）について議論してきた。$K > 2$の場合はどのように対処すべきか？

- 一対多（OVA）：K個のクラスのうち1つと残りの$(K - 1)$個のクラスを比較して、合計K個のSVMを当てはめる。$\hat{f}_k(x^*)$が最も大きいくなるようなクラスにx^*を分類する。

SVMs: 3つ以上のクラスがある場合は？

これまで、2つのクラスにおける分類の場合（$K = 2$）について議論してきた。$K > 2$の場合はどのように対処すべきか？

- 一対多（OVA）：K個のクラスのうち1つと残りの$(K - 1)$個のクラスを比較して、合計K個のSVMを当てはめる。$\hat{f}_k(x^*)$が最も大きいくなるようなクラスにx^*を分類する。

- 一対一（OVO）：$\binom{K}{2}$個のSVMを構成し、それぞれが2つのクラスを比較する。x^*はこれらのペアワイズの分類において最も多く割り当てられたクラスに割り当てる。

SVMs: 3つ以上のクラスがある場合は？

これまで、2つのクラスにおける分類の場合（$K = 2$）について議論してきた。$K > 2$の場合はどのように対処すべきか？

- 一対多（OVA）：K個のクラスのうち1つと残りの$(K - 1)$個のクラスを比較して、合計K個のSVMを当てはめる。$\hat{f}_k(x^*)$が最も大きいくなるようなクラスにx^*を分類する。

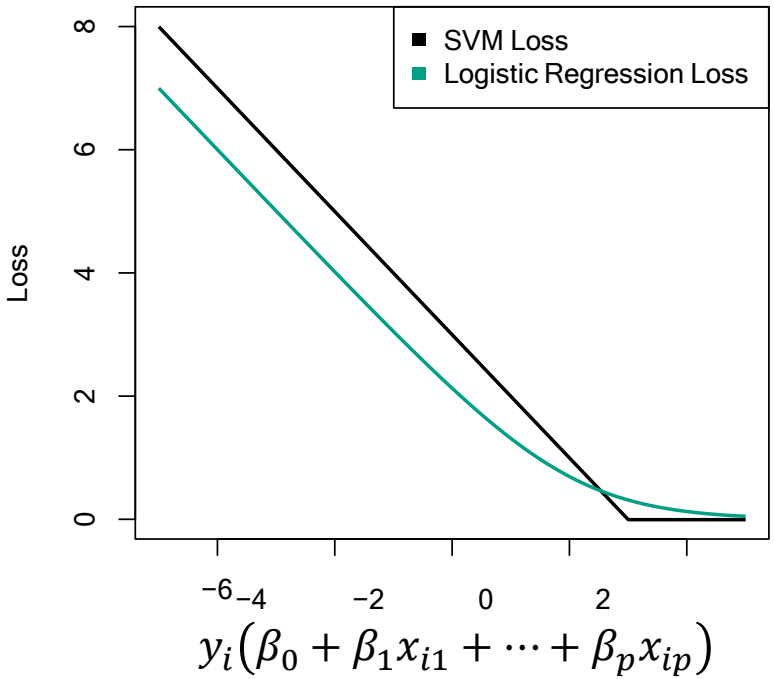
- 一対一（OVO）：$\binom{K}{2}$個のSVMを構成し、それぞれが2つのクラスを比較する。x^*はこれらのペアワイズの分類において最も多く割り当てられたクラスに割り当てる。

Kが大きくない場合は一対一を利用する。

サポートベクターとロジスティック回帰

サポートベクター分類器は次のように表現することができる。

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{minimize}} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$



この関数は「損失＋罰則」の形をとっている。
この損失は、ヒンジロスとして知られている。
ヒンジロスはロジスティック回帰の損失関数に似ている。

どれを使うか？：サポートベクター v.s. ロジスティック回帰

- 分離可能なデータに対して、SVMがロジスティック回帰や線形判別よりも優れた性能を示している。
- 分離不可能な場合は、ロジスティック回帰とSVMは類似した結果を与える。
- 確率を用いたい場合はロジスティック回帰を用いるべき。
- 非線形性がある場合は、カーネルSVMがよく用いられる。カーネル法をロジスティック回帰や線形判別に適用することが可能ではあるが、計算が複雑になってしまう。

第10章：深層学習
-Deep Learning-

- ニューラルネットワークの導入
- 畳込みニューラルネットワーク
- 文書分類
- 再帰型ニューラルネットワーク
- 時系列予測
- ニューラルネットワークのフィット
- 二重降下

第10章：深層学習
-Deep Learning-

ニューラルネットワークは1980年代に人気となった.
多くの成功や盛り上がりを見せ、NeurIPSやSnowbirdワークショップなど大きな会議も行われ始めた.

第10章：深層学習
-Deep Learning-

ニューラルネットワークは1980年代に人気となった.
多くの成功や盛り上がりを見せ、NeurIPSやSnowbirdワークショップ
など大きな会議も行われ始めた.

1990年代には、サポートベクターマシン、ランダムフォレスト、
ブースティングが流行り、ニューラルネットワークの人気は後退した.

第10章：深層学習
-Deep Learning-

ニューラルネットワークは1980年代に人気となった.
多くの成功や盛り上がりを見せ、NeurIPSやSnowbirdワークショップなど大きな会議も行われ始めた.

1990年代には、サポートベクターマシン、ランダムフォレスト、ブースティングが流行り、ニューラルネットワークの人気は後退した.

2010年前後に深層学習として再度現れ、2020年代にはかなり有力となり成功を収めるようになった.

成功の要因は、計算機能力の大きな改善や大量のトレーニングセットの収集、TensorflowやPyTorchといったソフトウェアが挙げられる.

第10章：深層学習
-Deep Learning-

ニューラルネットワークは1980年代に人気となった.
多くの成功や盛り上がりを見せ、NeurIPSやSnowbirdワークショップなど大きな会議も行われ始めた.

1990年代には、サポートベクターマシン、ランダムフォレスト、ブースティングが流行り、ニューラルネットワークの人気は後退した.

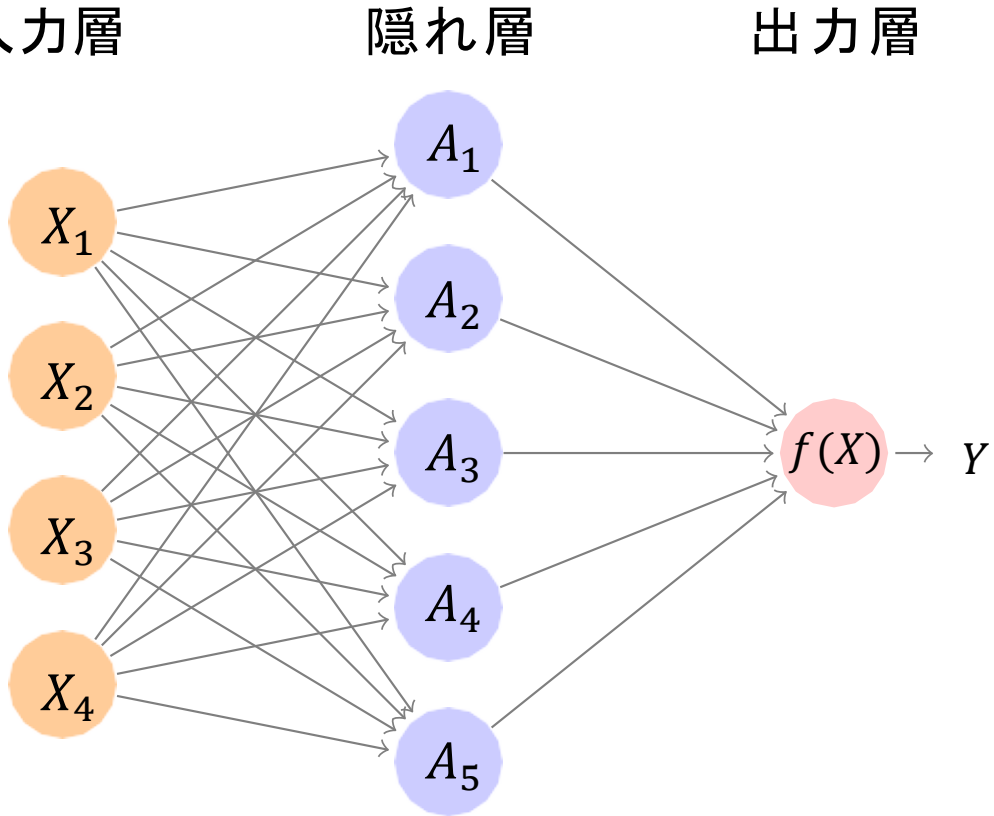
2010年前後に深層学習として再度現れ、2020年代にはかなり有力となり成功を収めるようになった.

成功の要因は、計算機能力の大きな改善や大量のトレーニングセットの収集、TensorflowやPyTorchといったソフトウェアが挙げられる.

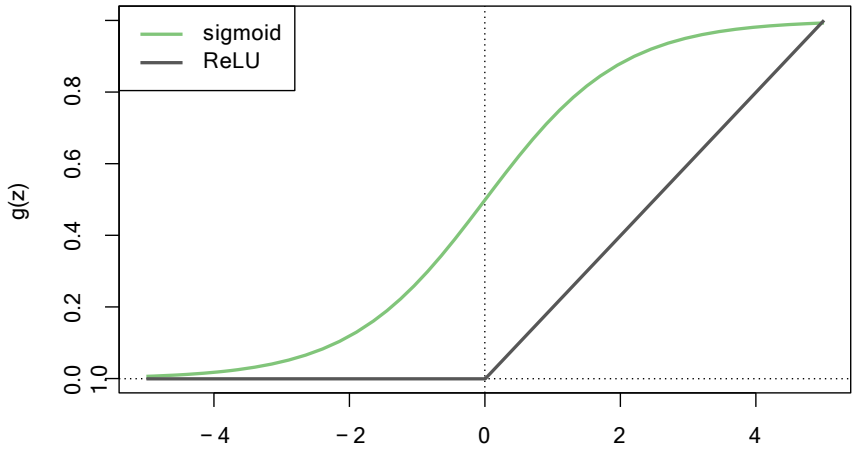
多くの功績が、Yann LeCun、Geoffrey Hinton、Yoshua Bengioとその学生らによるもので、2018年のACMのチューリング賞が与えられた.

隠れ層が1層のニューラルネットワーク

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X)$$
$$= \beta_0 + \sum_{k=1}^{K} \beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$$

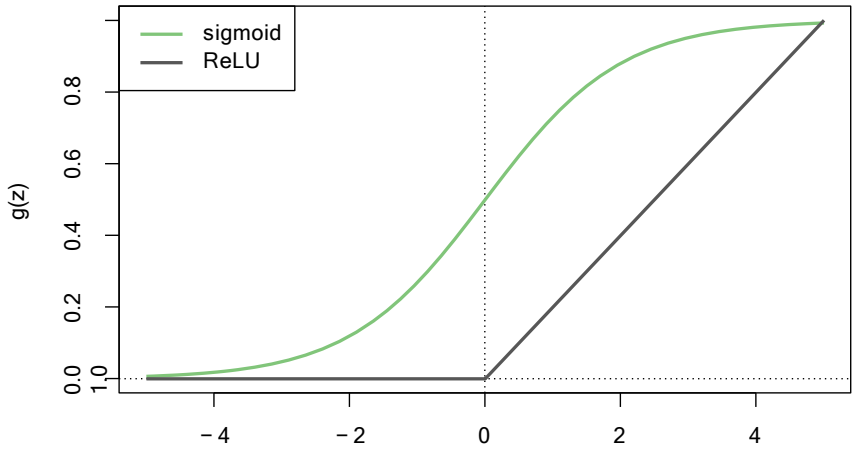


詳細



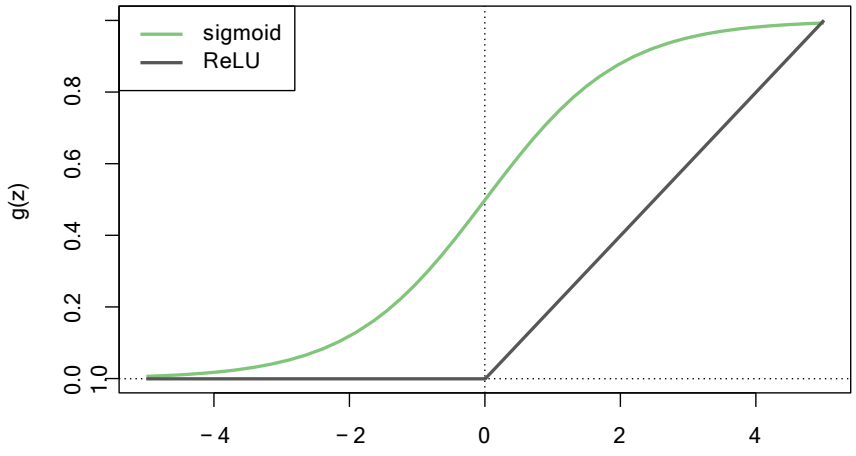
- $A_k = h_k(X) = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j\right)$ は隠れ層での活性化という.

詳細



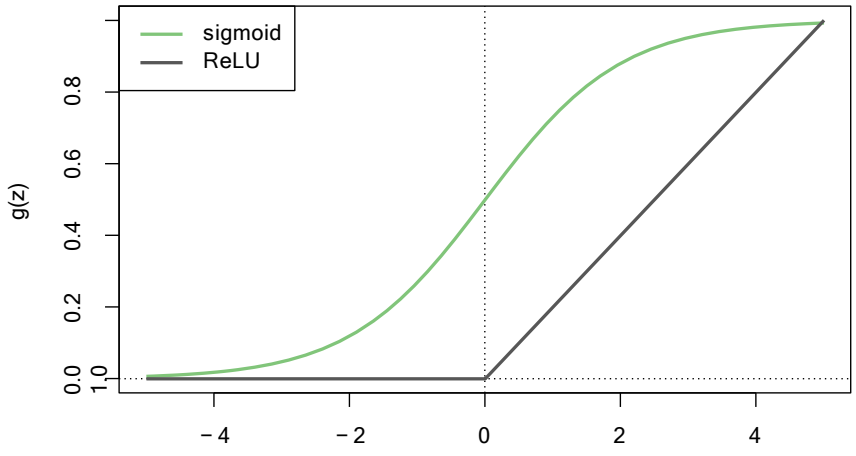
- $A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$ は隠れ層での活性化という.
- $g(z)$ は活性化関数と呼ばれる. 図で示されるようなシグモイドやReLU(正規化線形ユニット)がよく用いられる.

詳細



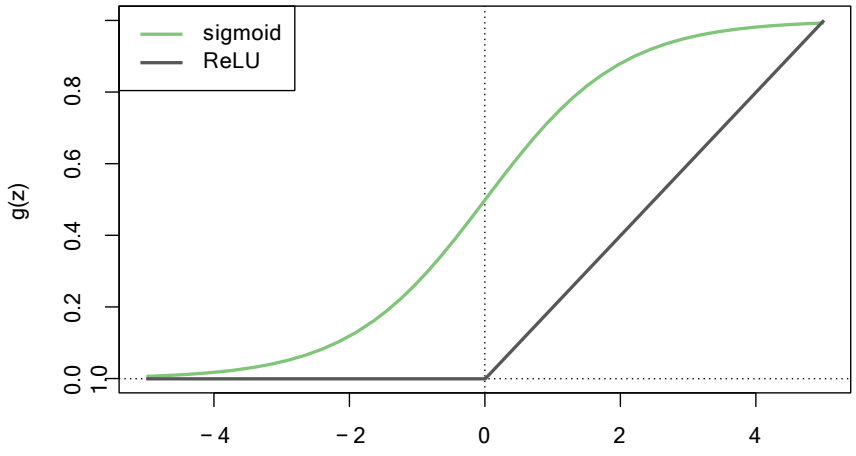
- $A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$ は隠れ層での活性化という.
- $g(z)$ は活性化関数と呼ばれる. 図で示されるようなシグモイドやReLU(正規化線形ユニット)がよく用いられる.
- 隠れ層の活性化関数は通常非線形のものを用いる. 線形の活性化関数を用いるとモデル全体として線形になってしまう.

詳細



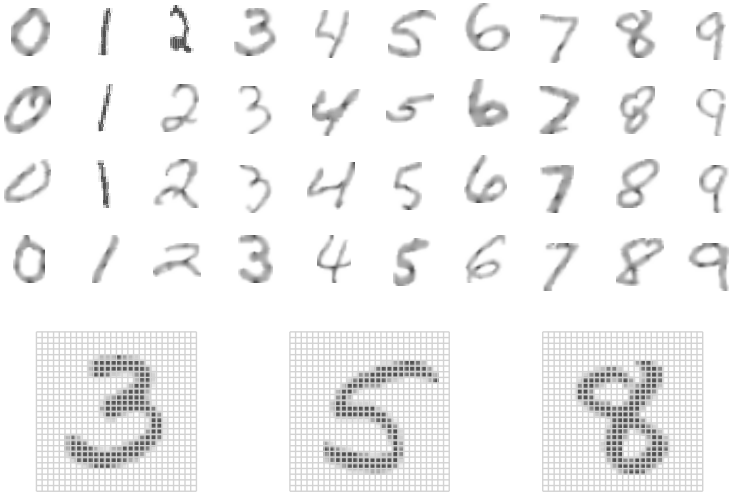
- $A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$ は隠れ層での活性化という.
- $g(z)$ は活性化関数と呼ばれる. 図で示されるようなシグモイドやReLU(正規化線形ユニット)がよく用いられる.
- 隠れ層の活性化関数は通常非線形のものを用いる. 線形の活性化関数を用いるとモデル全体として線形になってしまう.
- 活性化は、学習により得られた特徴のようなもので、特徴の線形結合の非線形変換という形となる.

詳細



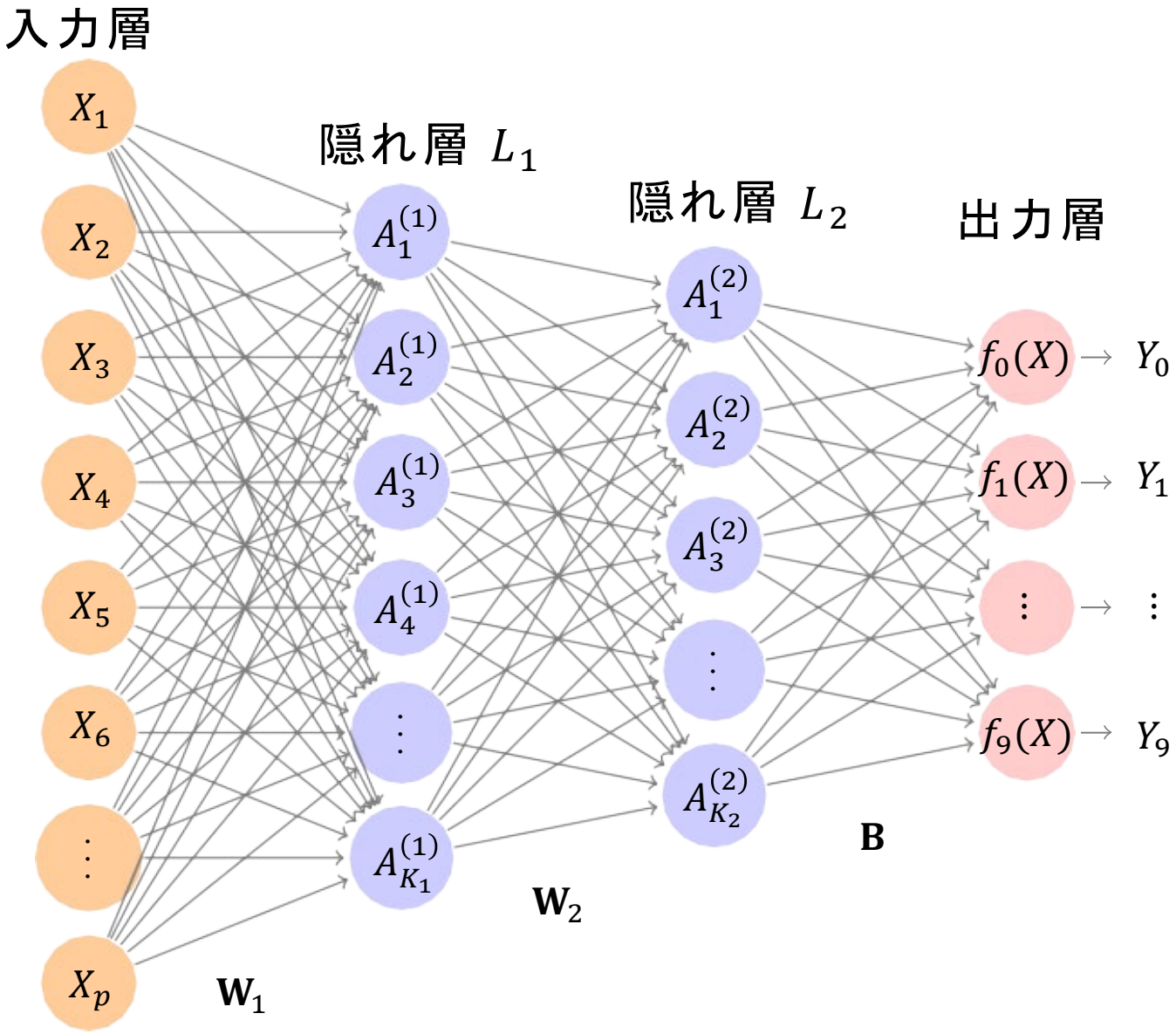
- $A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$ は隠れ層での活性化という.
- $g(z)$ は活性化関数と呼ばれる. 図で示されるようなシグモイドやReLU(正規化線形ユニット)がよく用いられる.
- 隠れ層の活性化関数は通常非線形のものを用いる. 線形の活性化関数を用いるとモデル全体として線形になってしまう.
- 活性化は、学習により得られた特徴のようなもので、特徴の線形結合の非線形変換という形となる.
- モデルのフィットは回帰のように、$\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2$を最小化することで行われる.

例: MNIST 数字



手書き数字
28×28グレースケールの画像
6万のトレーニング画像
1万のテスト画像
特徴は784ピクセルで
グレースケール値$\in (0, 255)$
ラベルは0–9の整数

- 目標:画像のクラスを予測するための分類器を構成する
- 第1層が256ユニット, 第2層が128ユニット, 出力層が10ユニットを持つ2層ネットワークを構成する
- (バイアスと呼ばれる)定数項と合わせて、(重みと呼ばれる)235,146個のパラメータがある


入力層
隠れ層 L_1
隠れ層 L_2
出力層
X_1
X_2
X_3
X_4
X_5
X_6
X_p
$A_1^{(1)}$
$A_2^{(1)}$
$A_3^{(1)}$
$A_4^{(1)}$
$A_{K_1}^{(1)}$
$A_1^{(2)}$
$A_2^{(2)}$
$A_3^{(2)}$
$A_{K_2}^{(2)}$
$f_0(X) \rightarrow Y_0$
$f_1(X) \rightarrow Y_1$
$f_9(X) \rightarrow Y_9$
$\mathbf{W}_1$
$\mathbf{W}_2$
$\mathbf{B}$

出力層の詳細

- $Z_m = \beta_{m0} + \sum_{\ell=1}^{K_2} \beta_{m\ell} A_\ell^{(2)}, m = 0,1, \ldots, 9$を第2層での10の活性化の線形結合とする
- 出力の活性化関数はソフトマックス関数とする

$$f_m(X) = \Pr(Y = m|X) = \frac{e^{Z_m}}{\sum_{\ell=0}^{9} e^{Z_\ell}}$$

出力層の詳細

- $Z_m = \beta_{m0} + \sum_{\ell=1}^{K_2} \beta_{m\ell} A_\ell^{(2)}, m = 0,1, \ldots, 9$を
  第2層での10の活性化の線形結合とする
- 出力の活性化関数はソフトマックス関数とする

$$f_m(X) = \Pr(Y = m|X) = \frac{e^{Z_m}}{\sum_{\ell=0}^{9} e^{Z_\ell}}$$

- 多項分布の負の対数尤度(クロスエントロピー)を最小化することで、
  モデルをフィットさせる

$$-\sum_{i=1}^{n}\sum_{m=0}^{9} y_{im} \log(f_m(x_i))$$

- y_{im}は、観測iの真のクラスがmのとき1、そうでなければ0という、
  ワンホットエンコード

結果

方法	テスト誤差
ニューラルネットワーク + リッジ正則化	2.3%
ニューラルネットワーク + ドロップアウト正則化	1.8%
多項ロジスティック回帰	7.2%
線形判別分析	12.7%

- ニューラルネットの初期の成功は1990年代にあった
- 多くのパラメータを持っているので、正則化は必須
- 正則化やフィッティングの詳細は後述

結果

方法	テスト誤差
ニューラルネットワーク + リッジ正則化	2.3%
ニューラルネットワーク + ドロップアウト正則化	1.8%
多項ロジスティック回帰	7.2%
線形判別分析	12.7%

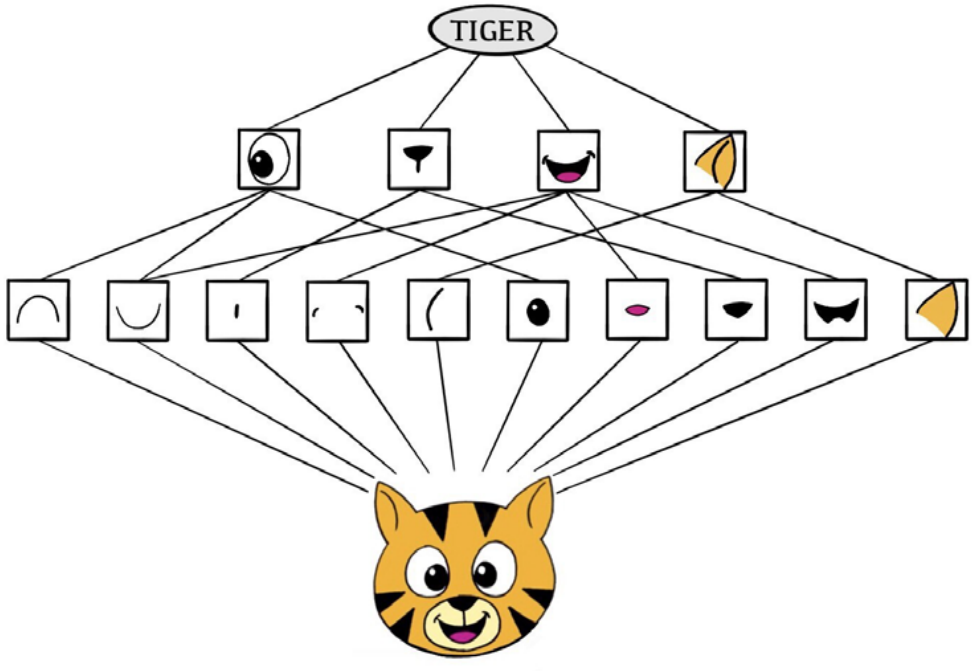
- ニューラルネットの初期の成功は1990年代にあった
- 多くのパラメータを持っているので、正則化は必須
- 正則化やフィッティングの詳細は後述
- あまりに多くの研究がなされていて、最も良い結果は $< 0.5\%$である!
- 人間による誤差は0.2%程度で、1万個のテスト画像の内20個程度の間違いである

畳込みニューラルネットワーク — CNN

- 画像の分類で大きな成功を果たした
- CIFAR100データセットからのサンプルを示している. 32×32 色の自然画像で100クラスを持っている
- 5万のトレーニング画像と1万のテスト画像がある

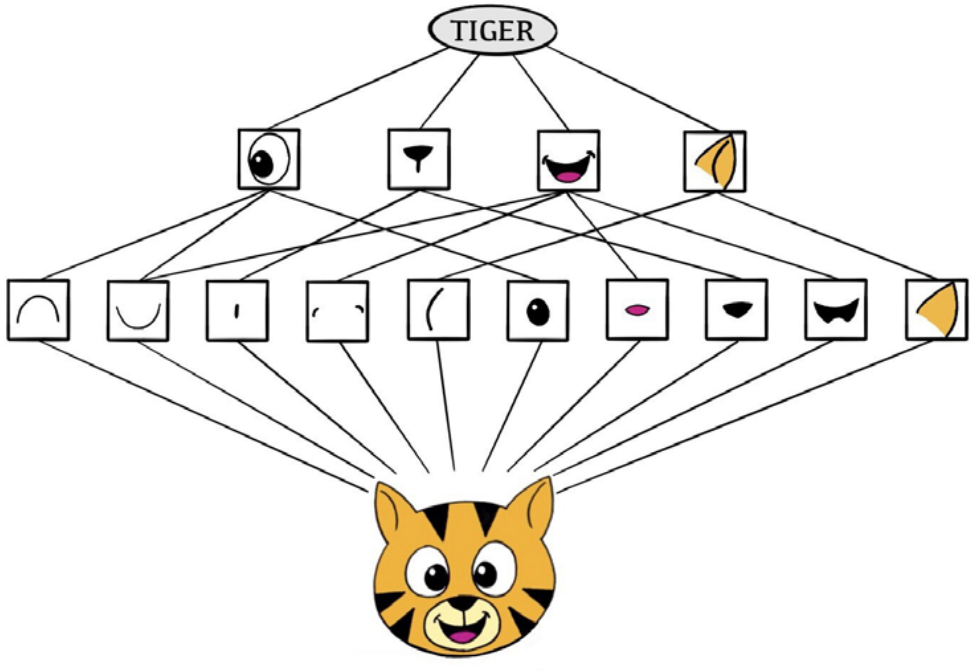
各画像は特徴マップの3次元配列である. 8ビットの数字の $32 \times 32 \times 3$ 配列. 最後の次元は赤、緑、青の3色のチャンネルを表している

CNNはどのような仕組みか



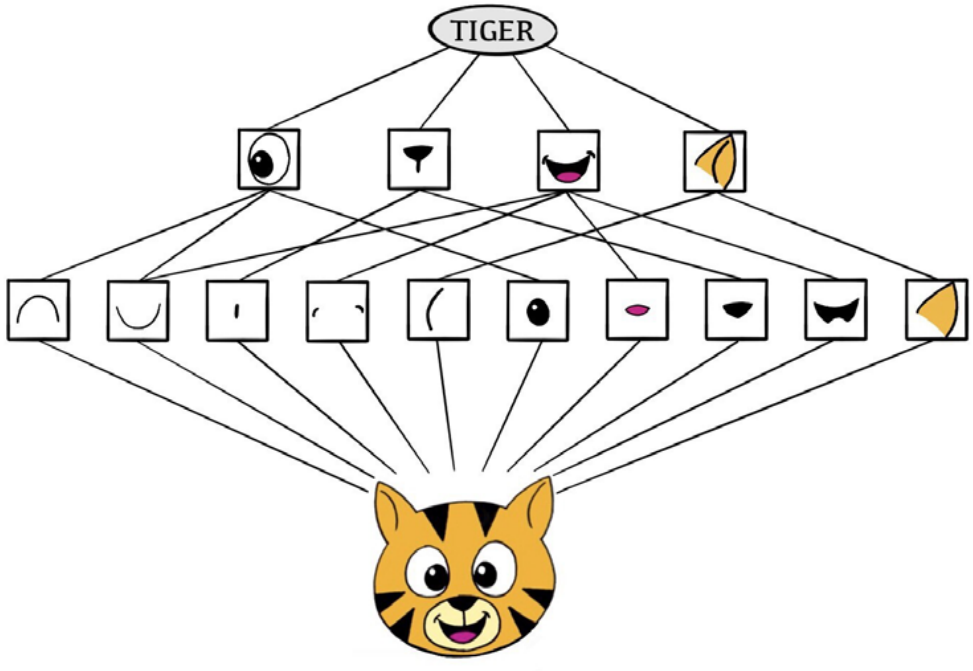
- CNNは階層的に画像を構成する

CNNはどのような仕組みか



- CNNは階層的に画像を構成する
- 境界や形が認識され、組み合わさる事でより複雑な形を形成し、最終的にターゲット画像を作り上げる

CNNはどのような仕組みか



- CNNは階層的に画像を構成する
- 境界や形が認識され、組み合わさる事でより複雑な形を形成し、最終的にターゲット画像を作り上げる
- この階層的な構成は畳み込み層とプーリング層を用いる事で可能となる

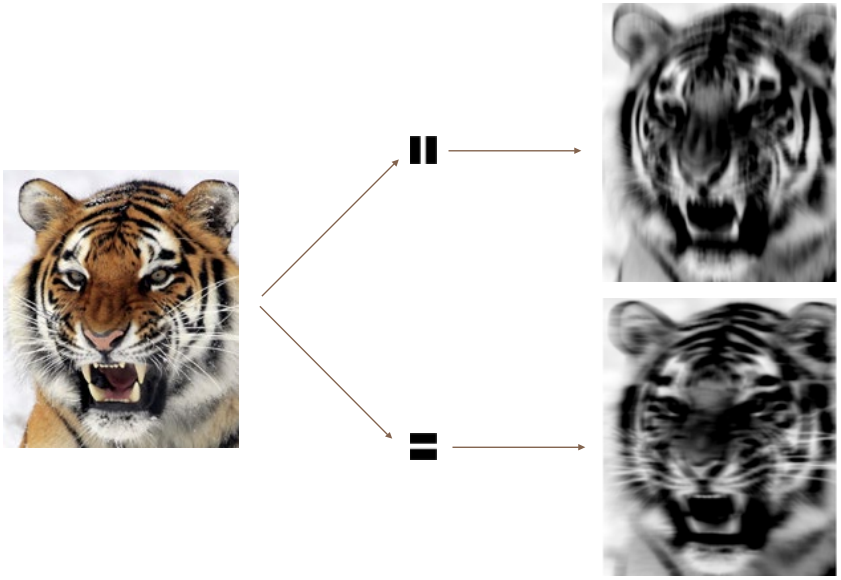
畳込みフィルタ

$$\text{入力画像}=\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix} \quad \text{畳込みフィルタ}=\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$$

$$\text{畳込み結果画像}=\begin{bmatrix} a\alpha+b\beta+d\gamma+e\delta & b\alpha+c\beta+e\gamma+f\delta \\ d\alpha+e\beta+g\gamma+h\delta & e\alpha+f\beta+h\gamma+i\delta \\ g\alpha+h\beta+j\gamma+k\delta & h\alpha+i\beta+k\gamma+l\delta \end{bmatrix}$$

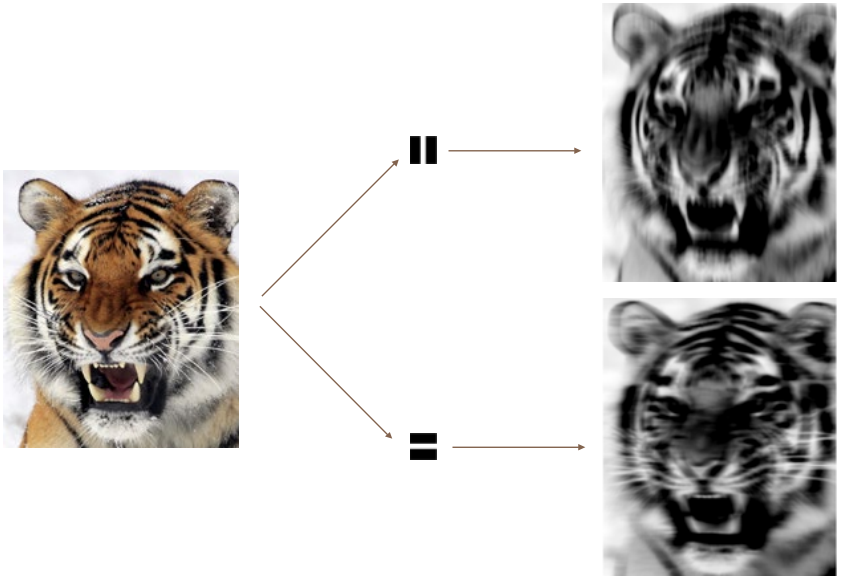
- フィルタはそれ自体が画像で、小さな形や境界などを表す
- 入力画像上をスライドさせて、一致度をスコアリングする
- スコアリングは上のようなドット積で行う
- もし入力画像のある部分がフィルタと類似していれば高いスコアとなり、類似していなければ低いスコアとなる
- フィルタは訓練の間に学習される

畳込みの例



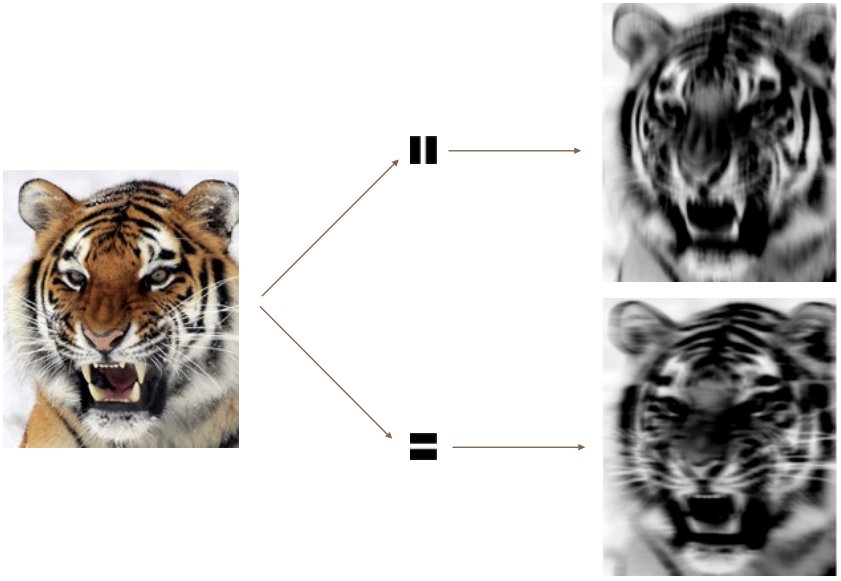
- フィルタを用いた畳込みのアイディアは、画像の異なる部分にある共通パターンを見つける事である

畳込みの例



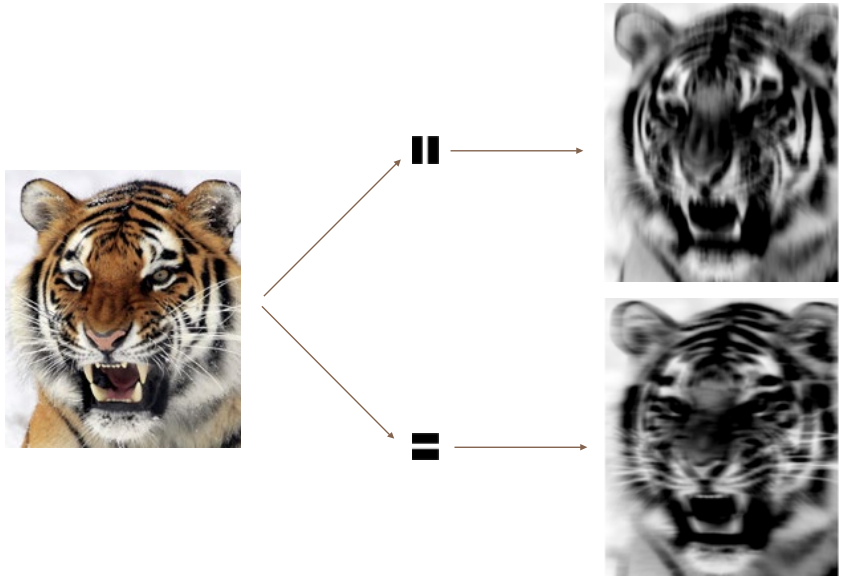
- フィルタを用いた畳込みのアイディアは、画像の異なる部分にある共通パターンを見つける事である
- ここでの2つのフィルタは垂直、水平方向の線を強調したものとなっている

畳込みの例



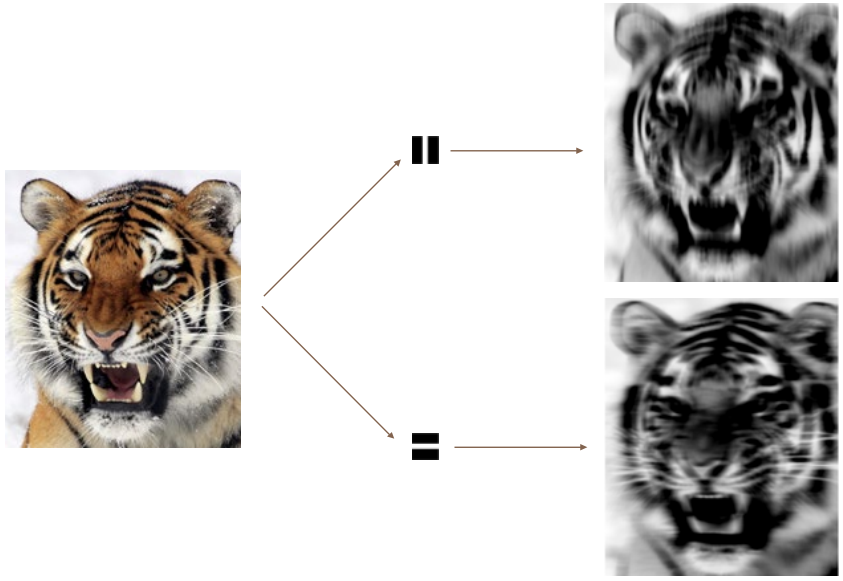
- フィルタを用いた畳込みのアイディアは、画像の異なる部分にある共通パターンを見つける事である
- ここでの2つのフィルタは垂直、水平方向の線を強調したものとなっている
- 畳込みの結果、新しい特徴マップが出来る

畳込みの例



- フィルタを用いた畳込みのアイディアは、画像の異なる部分にある共通パターンを見つける事である
- ここでの2つのフィルタは垂直、水平方向の線を強調したものとなっている
- 畳込みの結果、新しい特徴マップが出来る
- 画像は3色のチャンネルを持っているので、フィルタも3色のチャンネルを持つ. 1チャンネルに1つのフィルタがあり、ドット積の結果を足し合わせる

畳込みの例



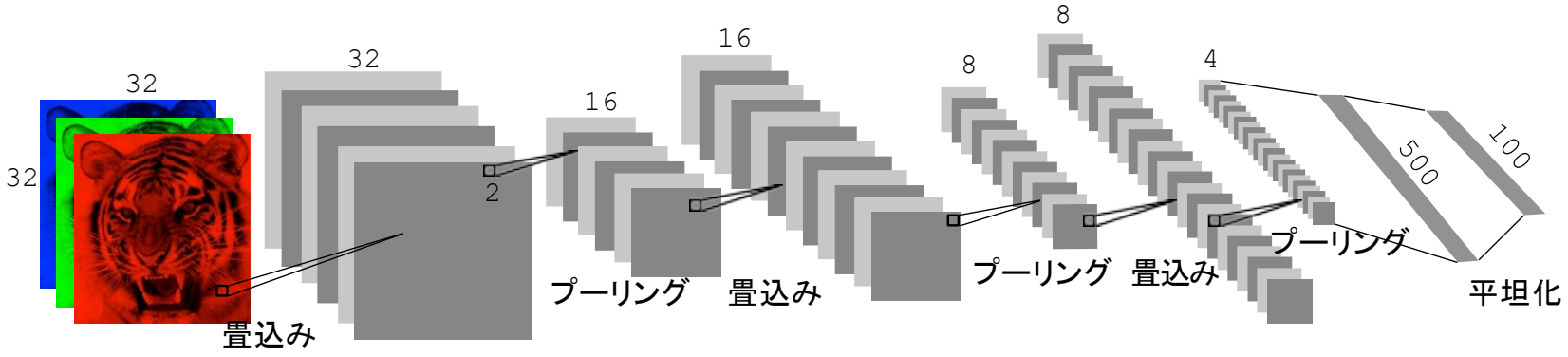
- フィルタを用いた畳込みのアイディアは、画像の異なる部分にある共通パターンを見つける事である
- ここでの2つのフィルタは垂直、水平方向の線を強調したものとなっている
- 畳込みの結果、新しい特徴マップが出来る
- 画像は3色のチャンネルを持っているので、フィルタも3色のチャンネルを持つ. 1チャンネルに1つのフィルタがあり、ドット積の結果を足し合わせる
- フィルタの重みはネットワークによって学習される

プーリング

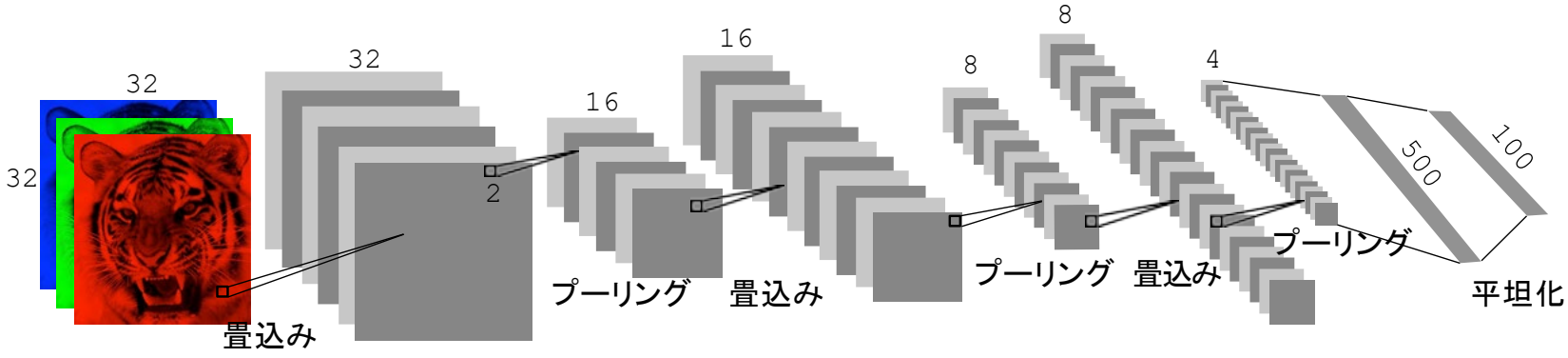
$$\text{最大プール} \begin{bmatrix} 1 & 2 & 5 & 3 \\ 3 & 0 & 1 & 2 \\ 2 & 1 & 3 & 4 \\ 1 & 1 & 2 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 3 & 5 \\ 2 & 4 \end{bmatrix}$$

- 重なりのない2×2行列毎にその最大値で置き換える
- これは特徴量を強調する
- 位置に関する不変性を許す
- 4つの要素だけによって次元削減している. 各次元について2つの要素となっている

CNNの構造

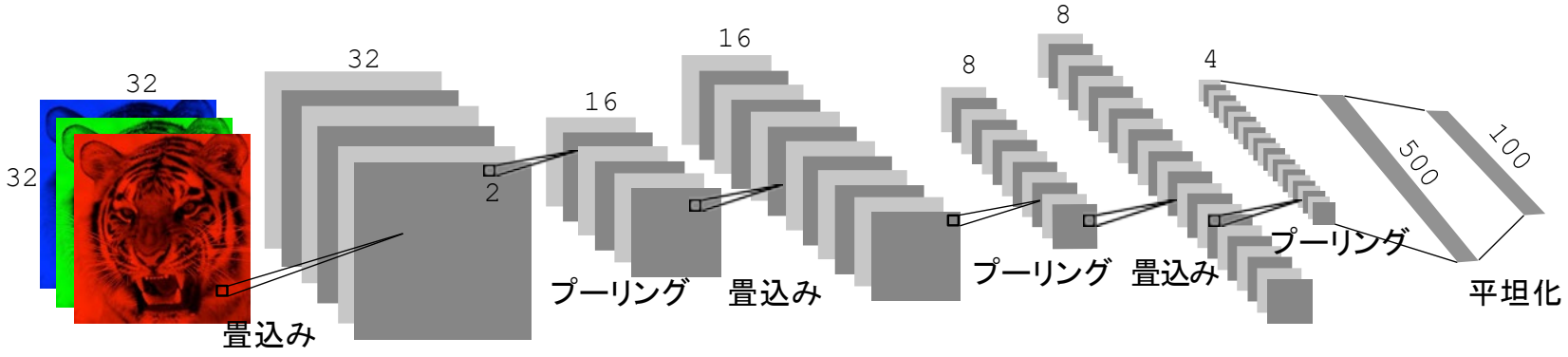


CNNの構造



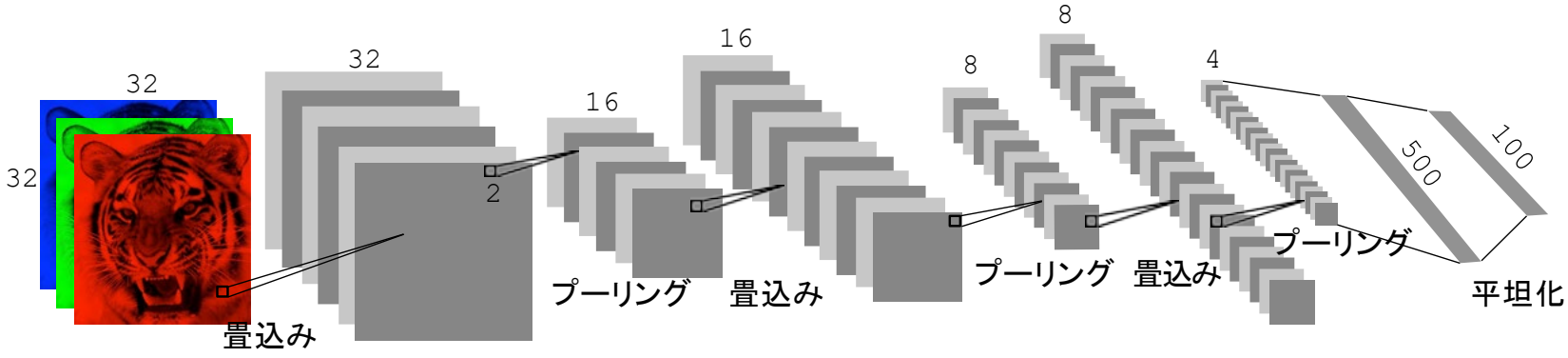
- 畳込み層 + プーリング層 をいくつも持つ

CNNの構造



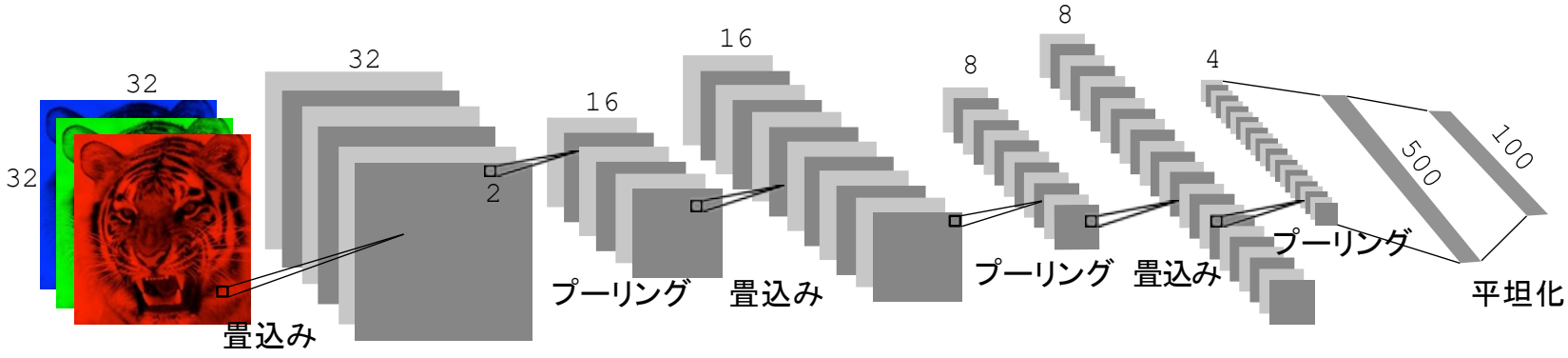
- 畳込み層 + プーリング層 をいくつも持つ
- フィルタは通常とても小さく、例えば各チャンネルは3 × 3

CNNの構造



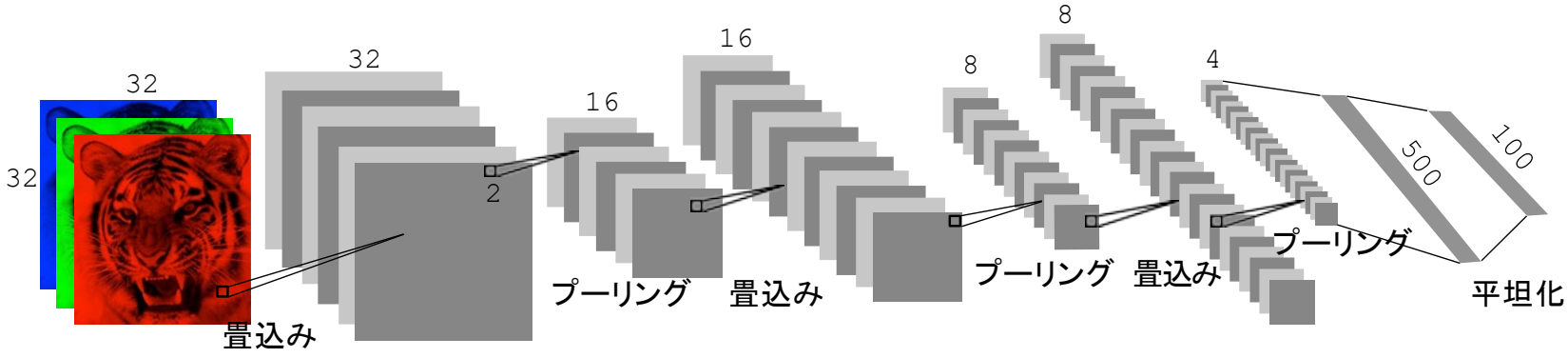
- 畳込み層 + プーリング層 をいくつも持つ
- フィルタは通常とても小さく、例えば各チャンネルは3 × 3
- 各フィルタは、畳込み層で、新たなチャンネルを作る

CNNの構造



- 畳込み層 + プーリング層 をいくつも持つ
- フィルタは通常とても小さく、例えば各チャンネルは3×3
- 各フィルタは、畳込み層で、新たなチャンネルを作る
- プーリングにより大きさが削減されるにつれ、フィルタやチャンネルの数は通常増加する

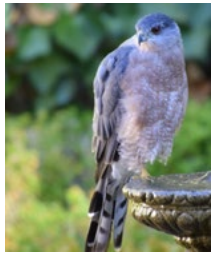
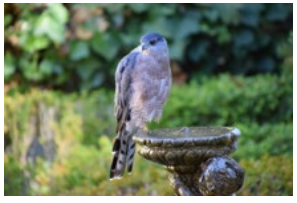
CNNの構造

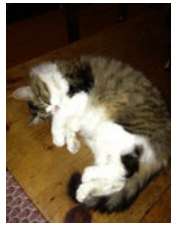


- 畳込み層 + プーリング層 をいくつも持つ
- フィルタは通常とても小さく、例えば各チャンネルは3 × 3
- 各フィルタは、畳込み層で、新たなチャンネルを作る
- プーリングにより大きさが削減されるにつれ、フィルタやチャンネルの数は通常増加する
- 層の数はとても大きくなり得る. 例えばimagenetの1000クラスの画像データベースを学習したresnet50は50層を持つ

学習済みネットワークを用いた画像の分類

学習済みネットワークを用いた画像の分類




フラミンゴ		クーパーハイタカ		クーパーハイタカ	
フラミンゴ	0.83	トビ	0.60	噴水	0.35
ヘラサギ	0.17	カラフトフクロウ	0.09	爪	0.12
シュバシコウ	0.00	コマドリ	0.06	フック	0.07

ラサアプソ		猫		ケープハタオリ	
チベタンテリア	0.56	オールドイングリッシュシープドッグ	0.82	キリハシ	0.28
ラサ	0.32	シーズー	0.04	コンゴウインコ	0.12
コッカースパニエル	0.03	ペルシャ猫	0.04	コマドリ	0.12

1000クラスのimagenetコーパスを用いて学習した50層のresnet50ネットワークを用いて、写真を分類する

文章分類: IMDB 映画レビュー

IMDBコーパスは、ユーザーが提供する多くの映画への評価からなる. それぞれはポジティブかネガティブとして感情をラベル付けされている. 次はネガティブな評価の冒頭である:

> This has to be one of the worst films of the 1990s. When my friends & I were watching this film (being the target audience it was aimed at) we just sat & watched the first half an hour with our jaws touching the floor at how bad it really was. The rest of the time, everyone else in the theater just started talking to each other, leaving or generally crying into their popcorn . . .
>
> (訳 : 1990年代の最悪の映画の1つであるに違いない. 友人とこの映画を見た時 、(私はこの映画が狙った視聴対象であっただろうが) 初めの30分を座って見てあまりの酷さに驚いてしまった. 残りの時間は映画館にいた全ての人がそれぞれに話し初め、席を立ったり後悔したりしていた. . .)

訓練セットとテストセットをそれぞれ25,000のレビューからなるように感情についてバランス良く分けた.

文章分類: IMDB 映画レビュー

IMDBコーパスは、ユーザーが提供する多くの映画への評価からなる. それぞれはポジティブかネガティブとして感情をラベル付けされている. 次はネガティブな評価の冒頭である:

> This has to be one of the worst films of the 1990s. When my friends & I were watching this film (being the target audience it was aimed at) we just sat & watched the first half an hour with our jaws touching the floor at how bad it really was. The rest of the time, everyone else in the theater just started talking to each other, leaving or generally crying into their popcorn . . .
>
> (訳 : 1990年代の最悪の映画の1つであるに違いない. 友人とこの映画を見た時 、(私はこの映画が狙った視聴対象であっただろうが) 初めの30分を座って見てあまりの酷さに驚いてしまった. 残りの時間は映画館にいた全ての人がそれぞれに話し初め、席を立ったり後悔したりしていた. . .)

訓練セットとテストセットをそれぞれ25,000のレビューからなるように感情についてバランス良く分けた.

レビューの感情を予測するための分類器を作りたい

特徴量化: Bag-of-Words

文章は異なる長さを持ち、単語の列で出来ている. どのようにして文章を特徴づけるような特徴Xを構成するか

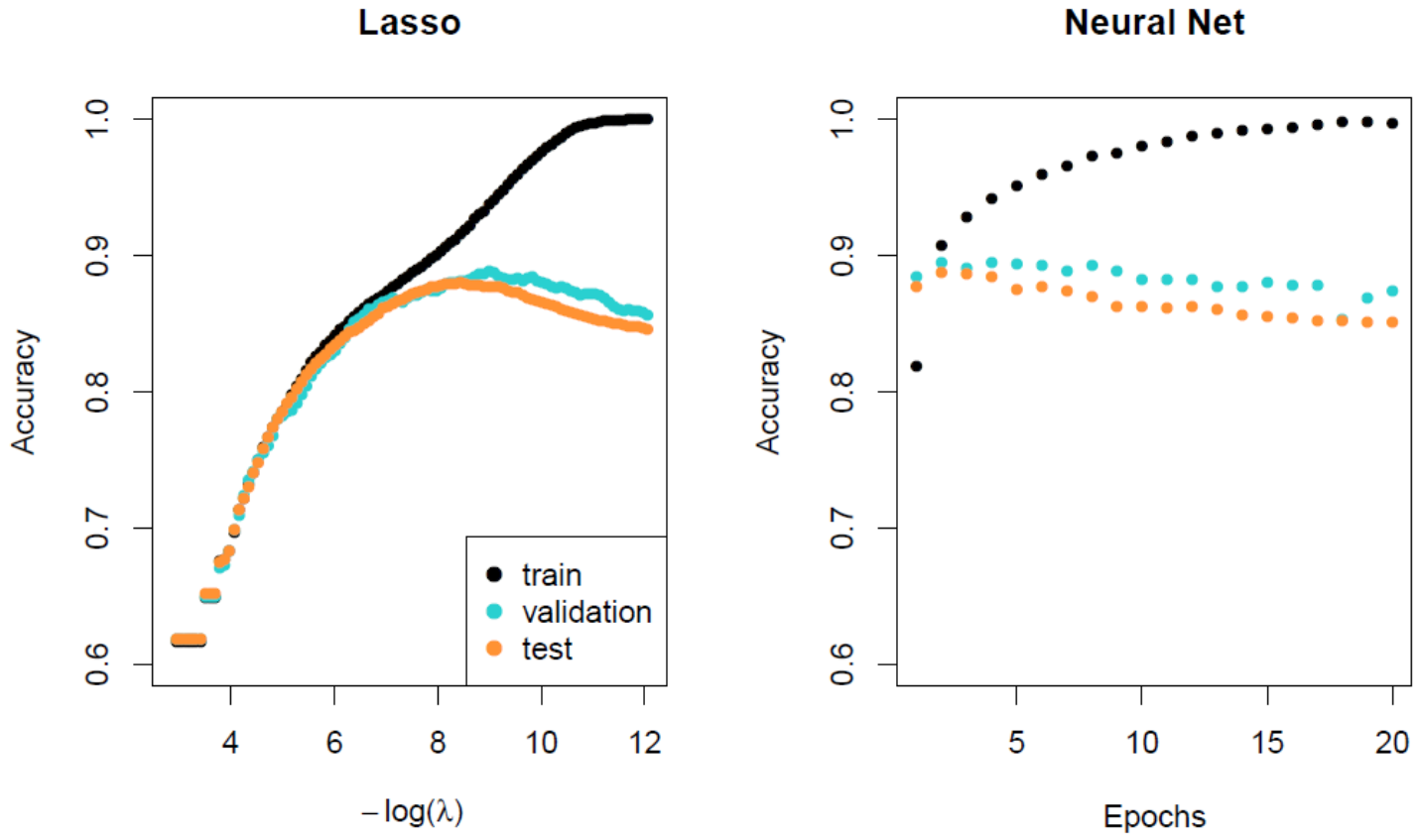
- 辞書から最頻の1万語を取ってくる
- それぞれの文章に対し、長さ$p = 1$万の2値ベクトルを作る. 各成分は単語が現れれば1とする
- n個の文章に対し、$n \times p$スパース特徴行列$\mathbf{X}$が出来る
- lassoロジスティック回帰モデルと隠れ層が2層のニューラルネットワークの比較を次スライドで行う. (ここでは畳み込みはなし)

特徴量化: Bag-of-Words

文章は異なる長さを持ち、単語の列で出来ている. どのようにして文章を特徴づけるような特徴Xを構成するか

- 辞書から最頻の1万語を取ってくる
- それぞれの文章に対し、長さ$p = 1$万の2値ベクトルを作る. 各成分は単語が現れれば1とする
- n個の文章に対し、$n \times p$スパース特徴行列$\mathbf{X}$が出来る
- lassoロジスティック回帰モデルと隠れ層が2層のニューラルネットワークの比較を次スライドで行う. (ここでは畳み込みはなし)
- Bag-of-wordsはユニグラムである. 代わりにバイグラム(隣接単語のペアの出現頻度)を用いる事もできる. 更に一般にm-グラムを用いる事もできる

Lasso 対 ニューラルネットワーク — IMDBレビュー



- このケースでは、シンプルなlassoロジスティック回帰モデルがニューラルネットワークと同等の性能
- glmnetがlassoモデルのフィットのために用いた. 行列**X**のスパース性を活用するために効果的である

再帰型ニューラルネットワーク

以下のようなデータがある

- 文章は単語の列で、相対的な位置関係に意味がある
- 天気データや金融指標のような時系列データ
- 録音データや音楽
- 医者のメモのような手書きのもの

RNNはデータのこのような順序の性質を取り入れ、過去の記憶を構成する

再帰型ニューラルネットワーク

以下のようなデータがある

- 文章は単語の列で、相対的な位置関係に意味がある
- 天気データや金融指標のような時系列データ
- 録音データや音楽
- 医者のメモのような手書きのもの

RNNはデータのこのような順序の性質を取り入れ、過去の記憶を構成する

- 各観測の特徴はベクトル$X = \{X_1, X_2, \ldots, X_L\}$の列である

再帰型ニューラルネットワーク

以下のようなデータがある

- 文章は単語の列で、相対的な位置関係に意味がある
- 天気データや金融指標のような時系列データ
- 録音データや音楽
- 医者のメモのような手書きのもの

RNNはデータのこのような順序の性質を取り入れ、過去の記憶を構成する

- 各観測の特徴はベクトル$X = \{X_1, X_2, \ldots, X_L\}$の列である
- ターゲットYはたいていここまでに現れた、感情のような一変量のものや多クラスのワンホットベクトルである

再帰型ニューラルネットワーク

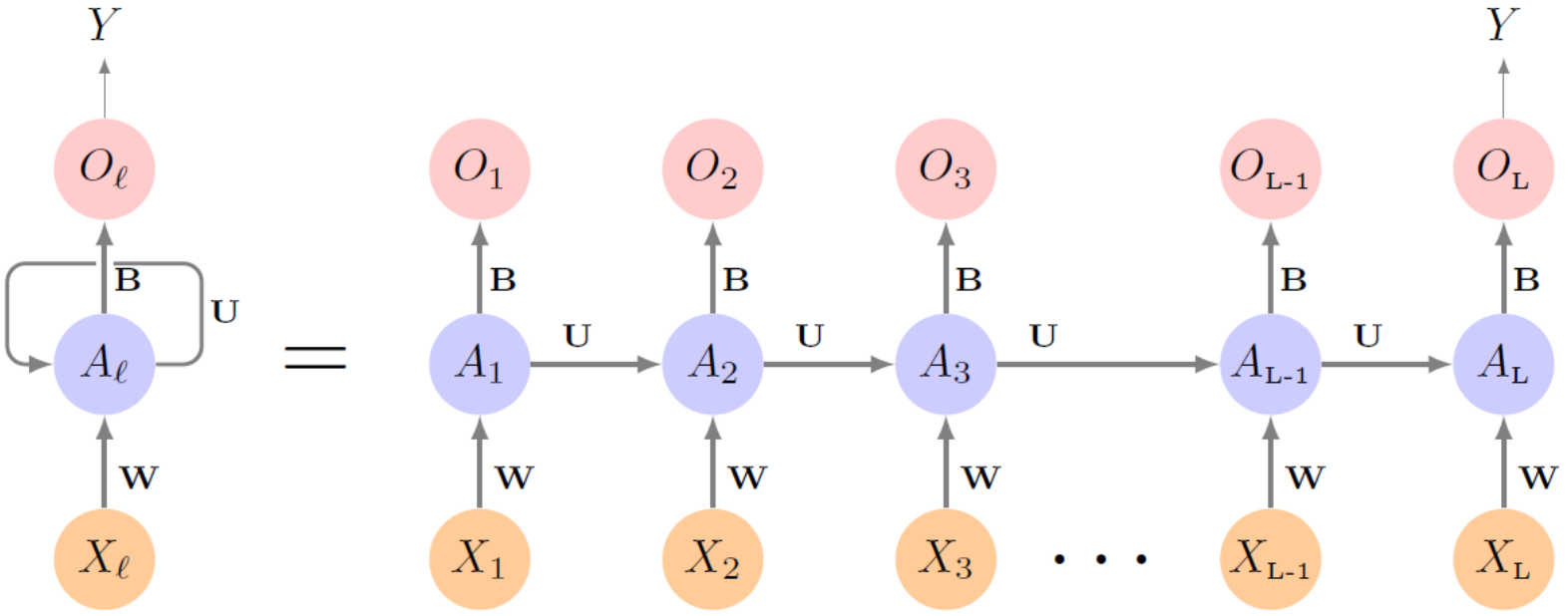
以下のようなデータがある

- 文章は単語の列で、相対的な位置関係に意味がある
- 天気データや金融指標のような時系列データ
- 録音データや音楽
- 医者のメモのような手書きのもの

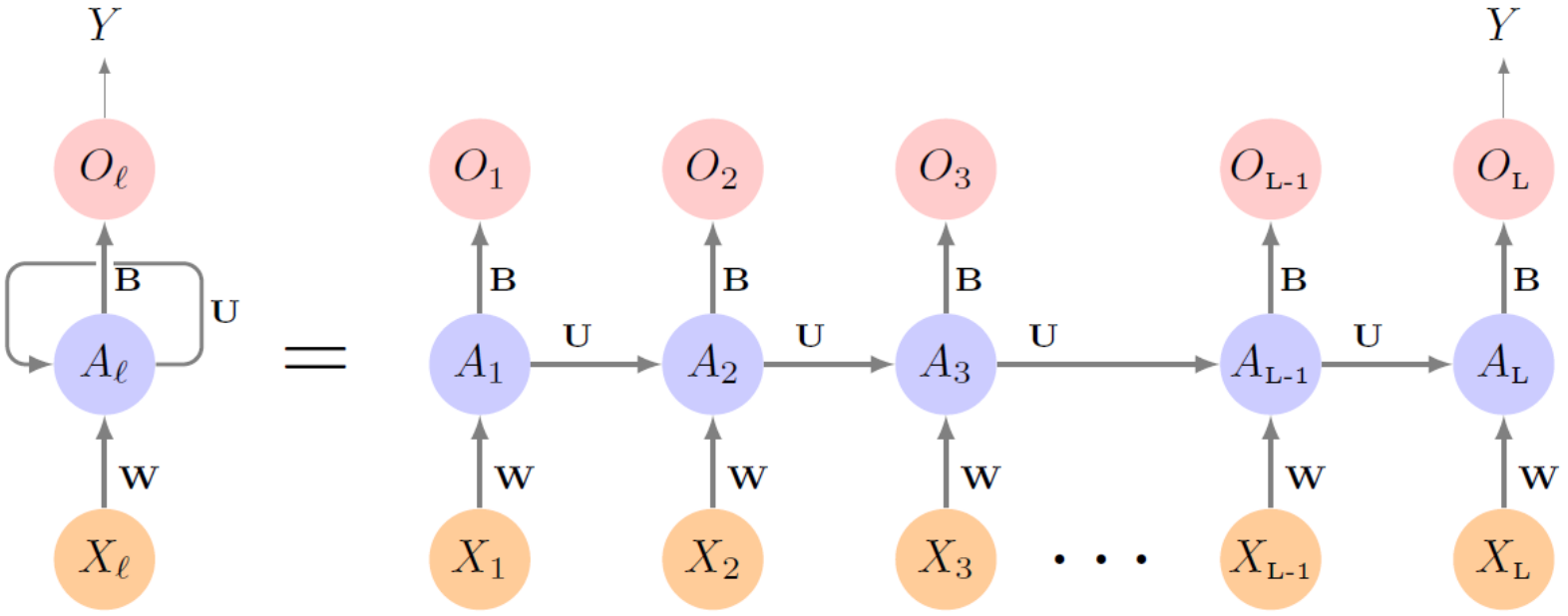
RNNはデータのこのような順序の性質を取り入れ、過去の記憶を構成する

- 各観測の特徴はベクトル$X = \{X_1, X_2, \ldots, X_L\}$の列である
- ターゲットYはたいていここまでに現れた、感情のような一変量のものや多クラスのワンホットベクトルである
- しかしYは、異なる言語による同じ文章のような、列でも良い

単純な再帰型ニューラルネットワークの構造

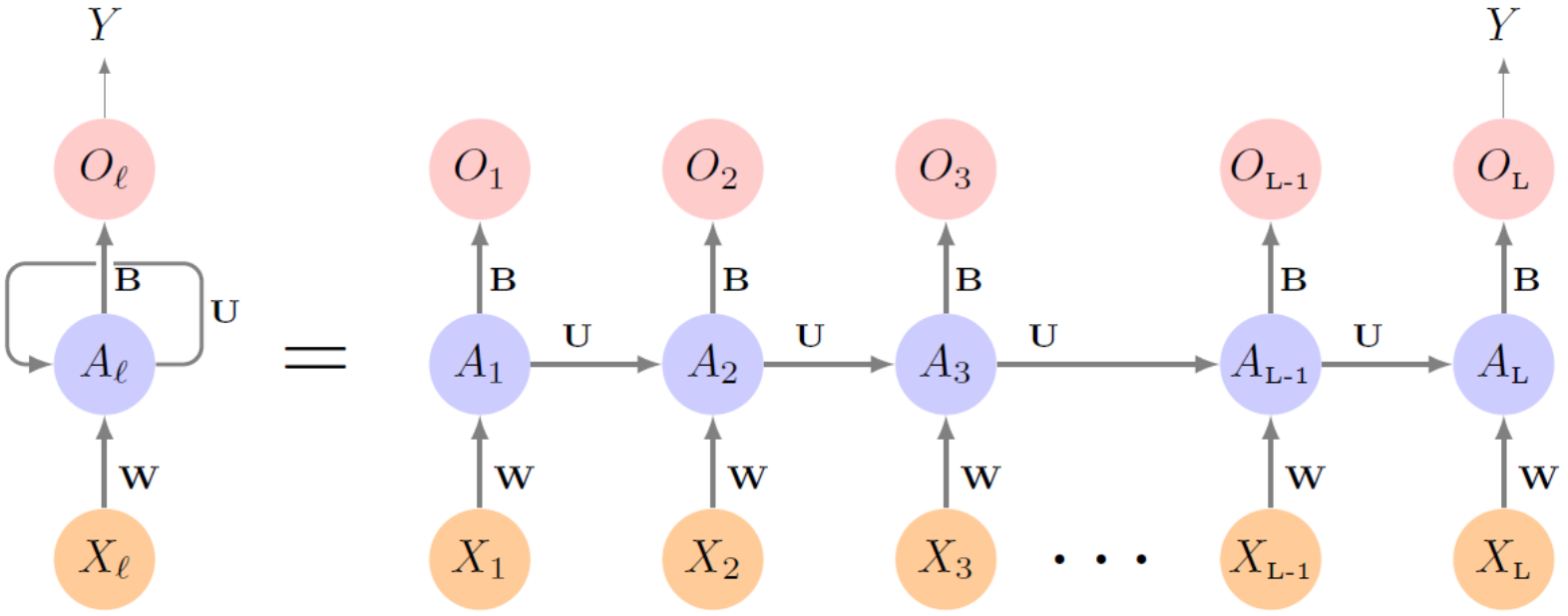


単純な再帰型ニューラルネットワークの構造



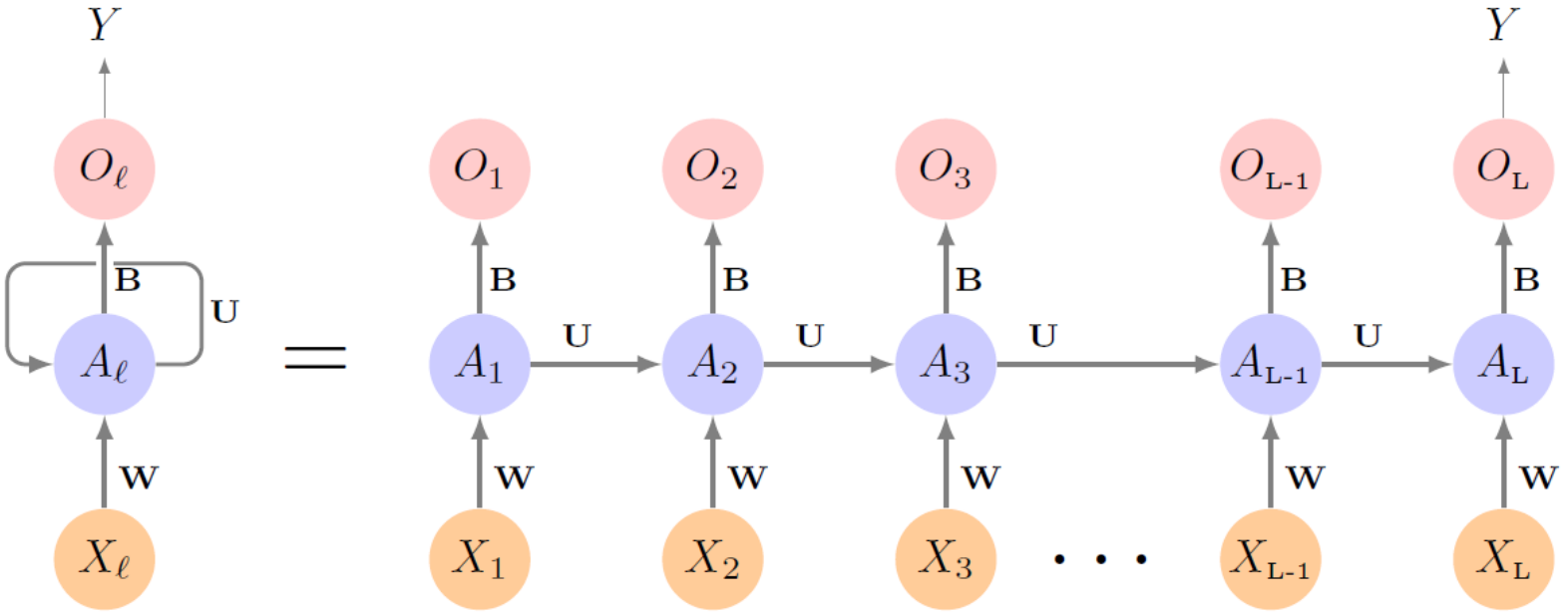
- 隠れ層はベクトルA_ℓの列で、$A_{\ell-1}$と入力X_ℓを受け取り、出力O_ℓを出す

単純な再帰型ニューラルネットワークの構造



- 隠れ層はベクトルA_ℓの列で、$A_{\ell-1}$と入力X_ℓを受け取り、出力O_ℓを出す
- 列の各ステップで同じ重み**W**, **U**, **B**を用いる. なので、再帰という

単純な再帰型ニューラルネットワークの構造



- 隠れ層はベクトルA_ℓの列で、$A_{\ell-1}$と入力X_ℓを受け取り、出力O_ℓを出す
- 列の各ステップで同じ重み**W**, **U**, **B**を用いる. なので、再帰という
- A_ℓの列は、X_ℓが処理される度に更新される、応答への発展的なモデルを表現している

RNNの詳細

$X_\ell = (X_{\ell 1}, X_{\ell 2}, \ldots, X_{\ell p}), \mathrm{A}_\ell = (\mathrm{A}_{\ell 1}, \mathrm{A}_{\ell 2}, \ldots, A_{\ell K})$とする. このとき、隠れユニット$A_\ell$の$k$番目の成分は

$$A_{\ell k} = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_{\ell j} + \sum_{s=1}^{K} u_{ks} A_{\ell-1,s}\right)$$

$$O_\ell = \beta_0 + \sum_{k=1}^{K} \beta_k A_{\ell k}$$

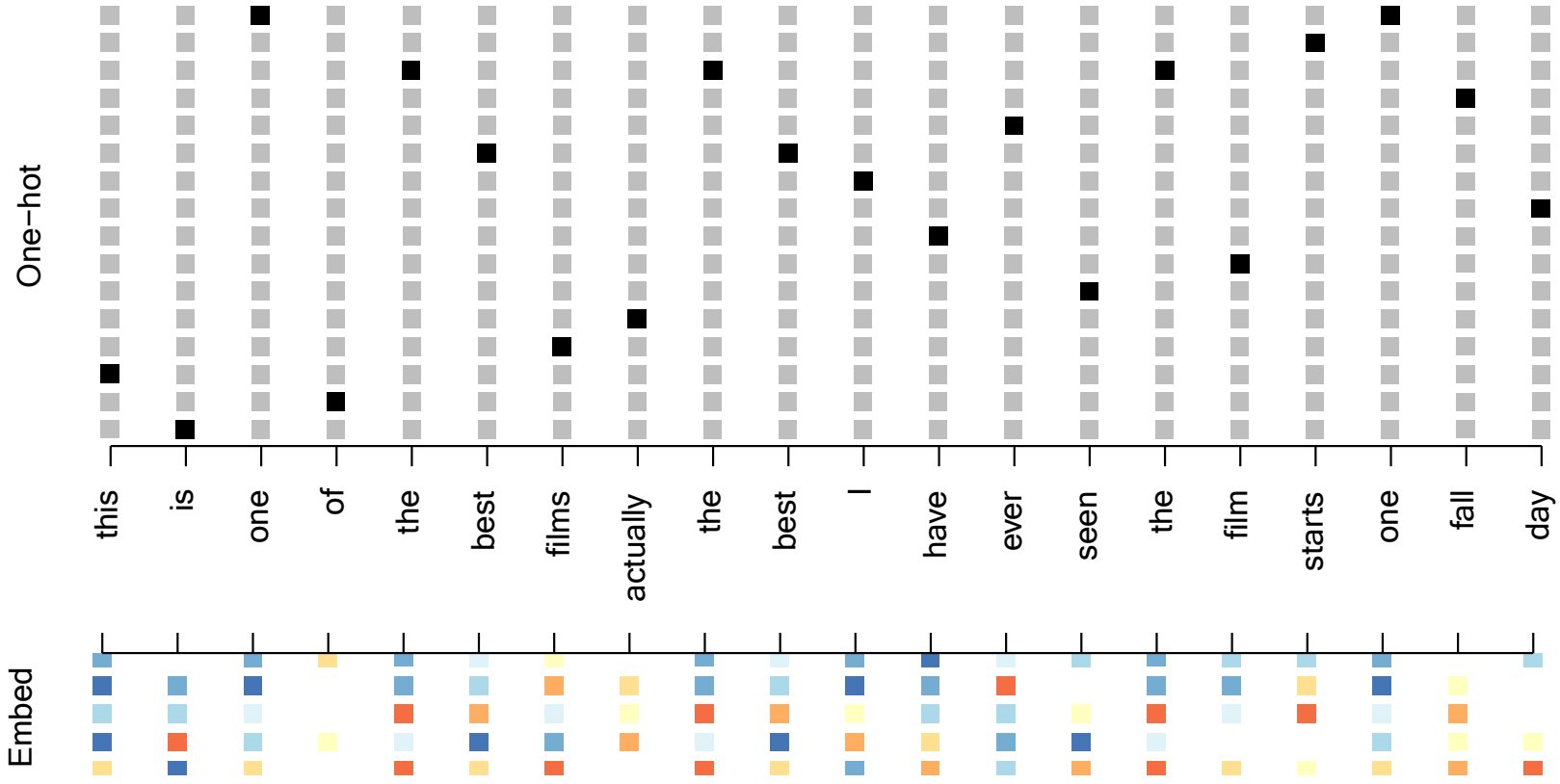
興味の対象は最後のユニットのO_Lの予測だけであることがよくある.
二乗損失に対し、n個の列/応答のペアに対し、次を最小化する.

$$\sum_{i=1}^{n} (y_i - o_{iL})^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{iL_j} + \sum_{s=1}^{K} u_{ks} a_{i,L-1,s}\right)\right)\right)^2$$

RNNとIMDBレビュー

- 文章の特徴は単語の列$\{\mathcal{W}_\ell\}_1^L$である. たいてい同じ長さLに対して打ち切ったり、膨らませたりする
- 各単語$\mathcal{W}_\ell$は長さ1万のワンホットエンコードの2値ベクトルX_ℓ (ダミー変数) 、辞書に現れる単語の位置だけ1で残りは0で表される
- この結果、かなりスパースな特徴表現となり、上手く行かない
- 代わりに低次元の学習済みの単語の埋め込み行列$\mathbf{E}(m \times 1$万, 次スライド)を用いる
- これによって、1万の長さの2値特徴ベクトルを$m(\ll 1$万$)$次元の実特徴ベクトルに削減する (mは例えば数百)

単語の埋め込み



this is one of the best films actually the best I have ever seen the film starts one fall day $\cdots$.

埋め込みは主成分に似た方法で、大きな文章のコーパスを用いて学習される. word2vecやGloVeはとても人気である

IMDBレビューへのRNN

- 多くの研究の結果を経ても、残念ながらこの精度は76%である

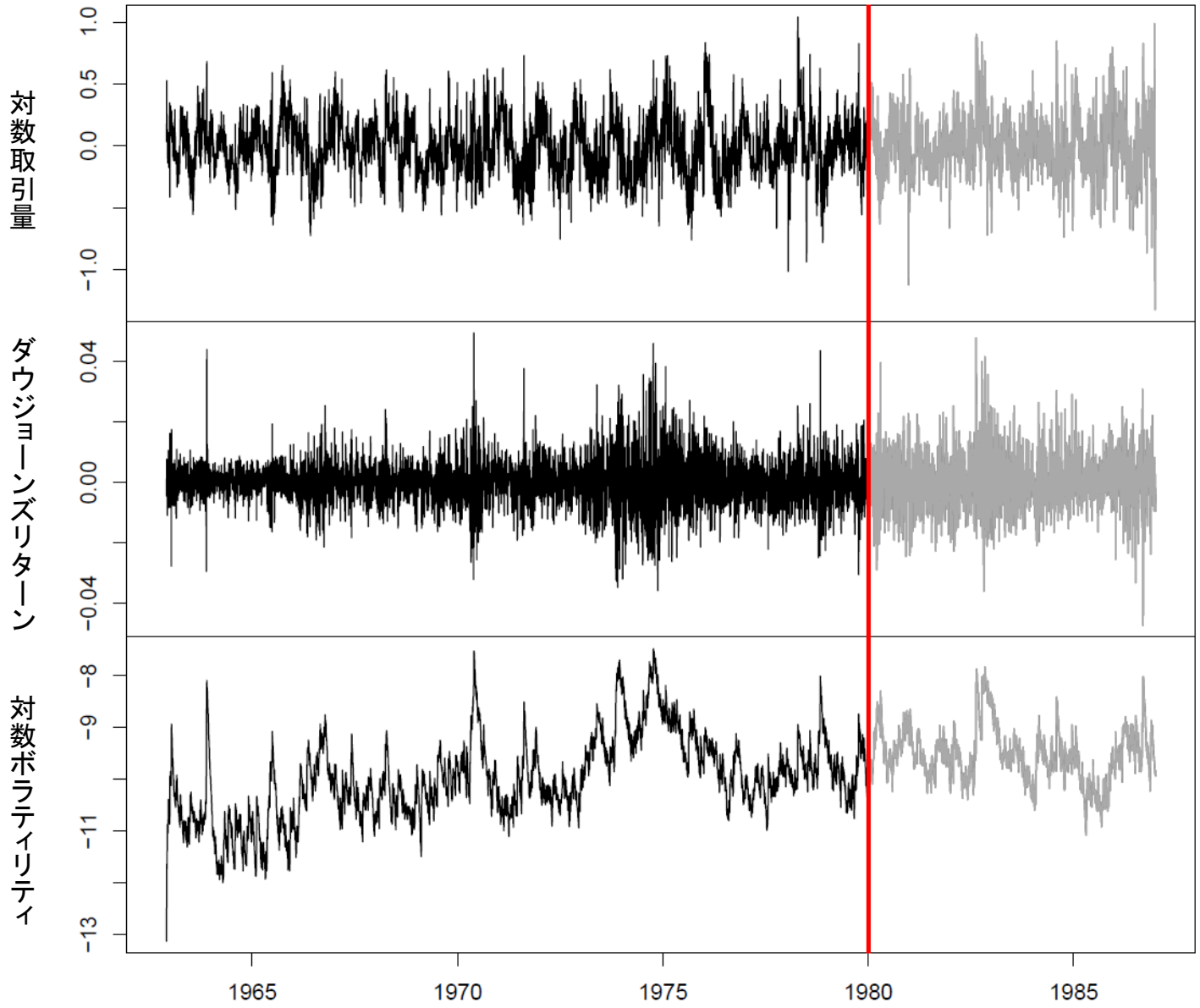
IMDBレビューへのRNN

- 多くの研究の結果を経ても、残念ながらこの精度は76%である
- 先に示したものより少し風代わりなRNN、長期記憶と短期記憶をもつLSTM. ここでA_ℓは(短期記憶である)$A_{\ell-1}$からの入力を受け取り、(長期記憶である)時間をさらに遡ったものからの入力も受け取る. 結果87%の精度でglmnetによる88%にわずかに及ばない程度となる

IMDBレビューへのRNN

- 多くの研究の結果を経ても、残念ながらこの精度は76%である
- 先に示したものより少し風代わりなRNN、長期記憶と短期記憶をもつLSTM. ここでA_ℓは(短期記憶である)$A_{\ell-1}$からの入力を受け取り、(長期記憶である)時間をさらに遡ったものからの入力も受け取る. 結果87%の精度でglmnetによる88%にわずかに及ばない程度となる
- このデータは新たなRNNの構造のベンチマークとして用いられ、2020年時点での最良の結果は95%程度である. セクション10.5.1でリーダーボードを示す

時系列予測


対数取引量
1.0
0.5
0.0
-1.0
ダウジョーンズリターン
0.04
0.00
-0.04
対数ボラティリティ
-8
-9
-11
-13
1965
1970
1975
1980
1985

ニューヨーク証券取引データ

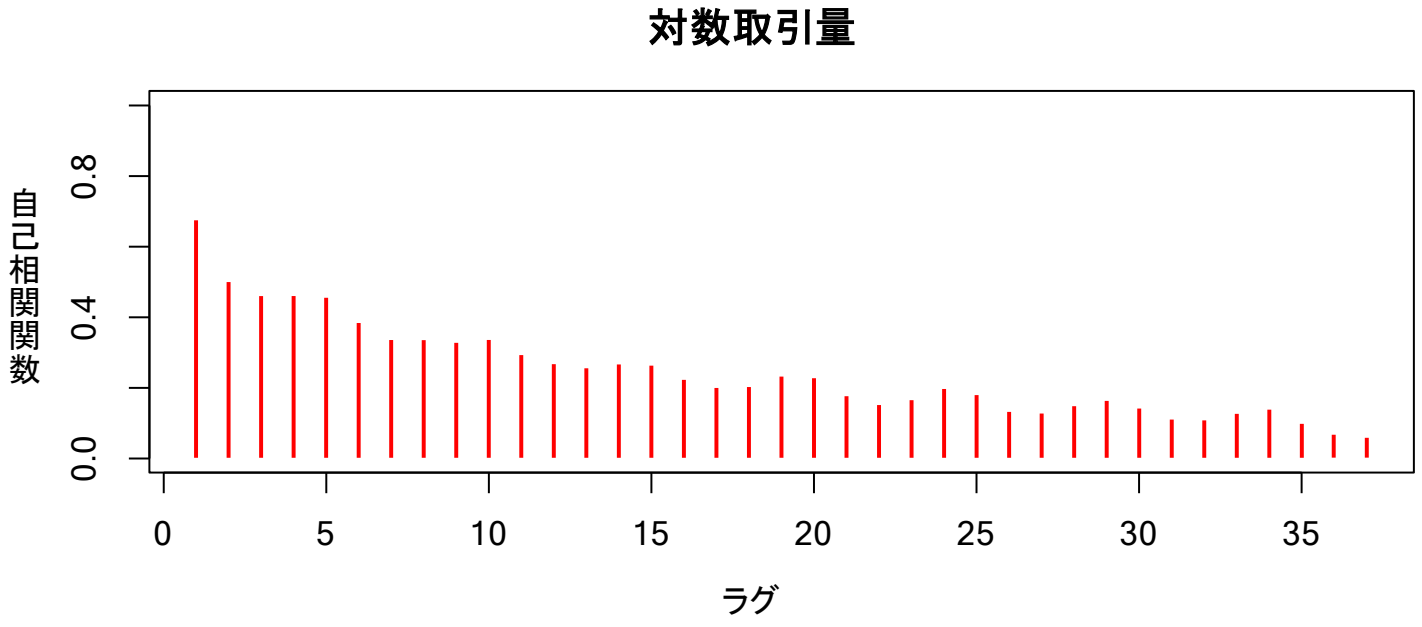
前スライドのように3つの1962年12月3日から1986年12月31日までの日次データ (6051取引日)

- 対数取引量. その日に取引された市場の全ての株式の、過去100日の移動平均のものに対する割合を対数スケールで表したもの
- ダウジョーンズリターン. 連続する取引日でのダウ平均株価の対数の差
- 対数ボラティリティ. 価格の日次変動の絶対値に基づく

目標: 今日までの観測から、明日の対数取引量を予測する事.
ダウジョーンズリターンと対数ボラティリティについても同様

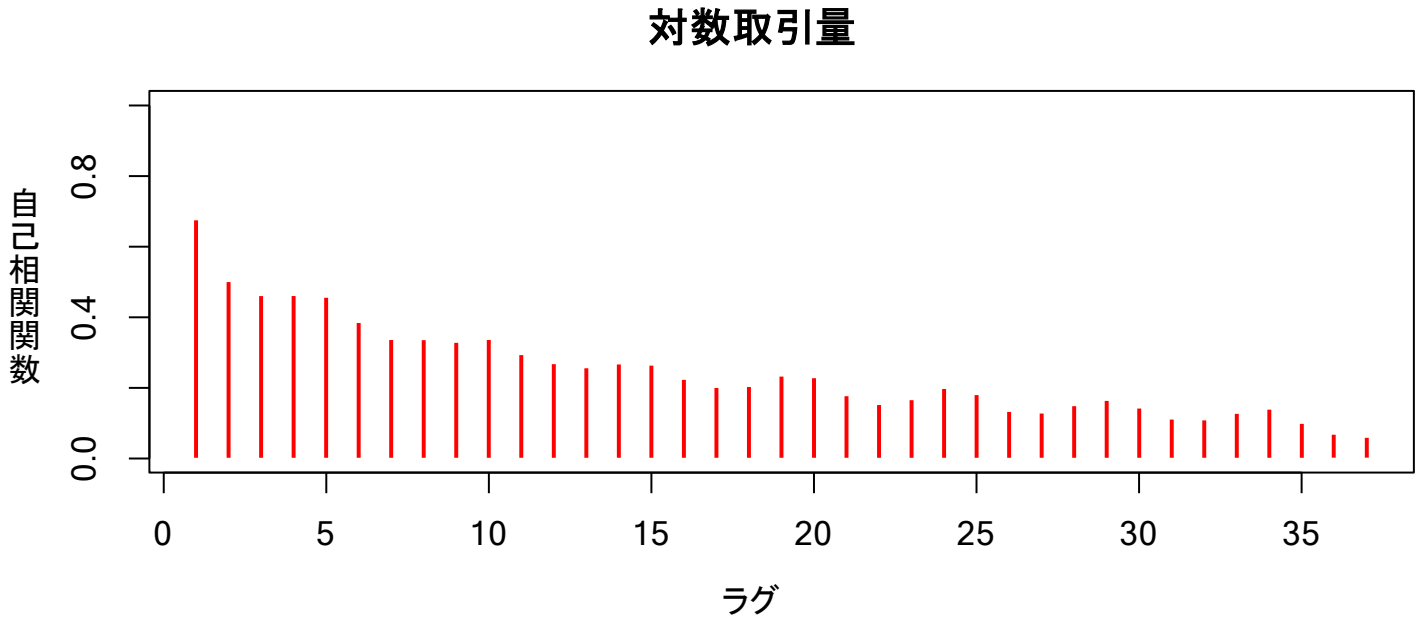
これらのデータは、LeBaron and Weigend (1998) *IEEE Transactions on Neural Networks*, 9(1): 213–220 による.

自己相関



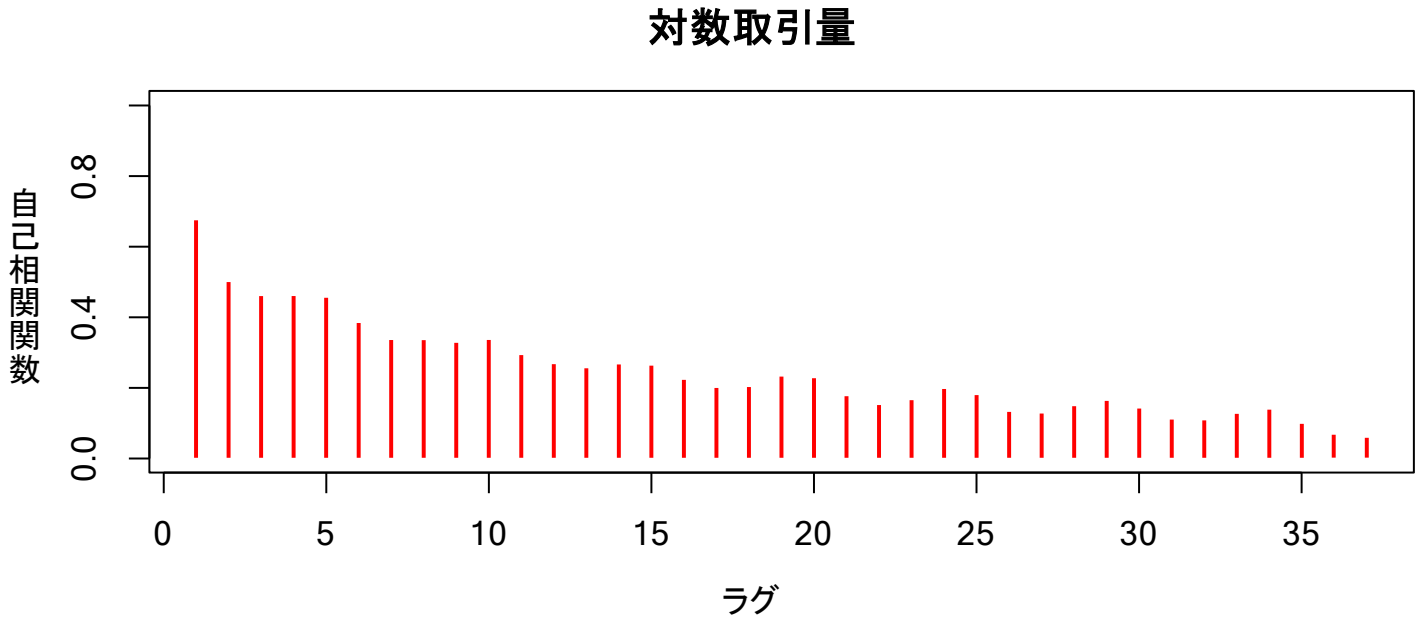
- ラグℓの自己相関はℓ取引日離れた$(v_t, v_{t-\ell})$の相関

自己相関



- ラグℓの自己相関はℓ取引日離れた$(v_t, v_{t-\ell})$の相関
- この大きな値の相関から、過去の値が将来の予測をするために役立つと確信できる

自己相関



- ラグℓの自己相関はℓ取引日離れた$(v_t, v_{t-\ell})$の相関
- この大きな値の相関から、過去の値が将来の予測をするために役立つと確信できる
- これは不思議な予測問題で、応答v_tが特徴$v_{t-\ell}$でもある

RNNによる予測

データの列を1つだけ持っている. どのようにしてRNNを作るか.

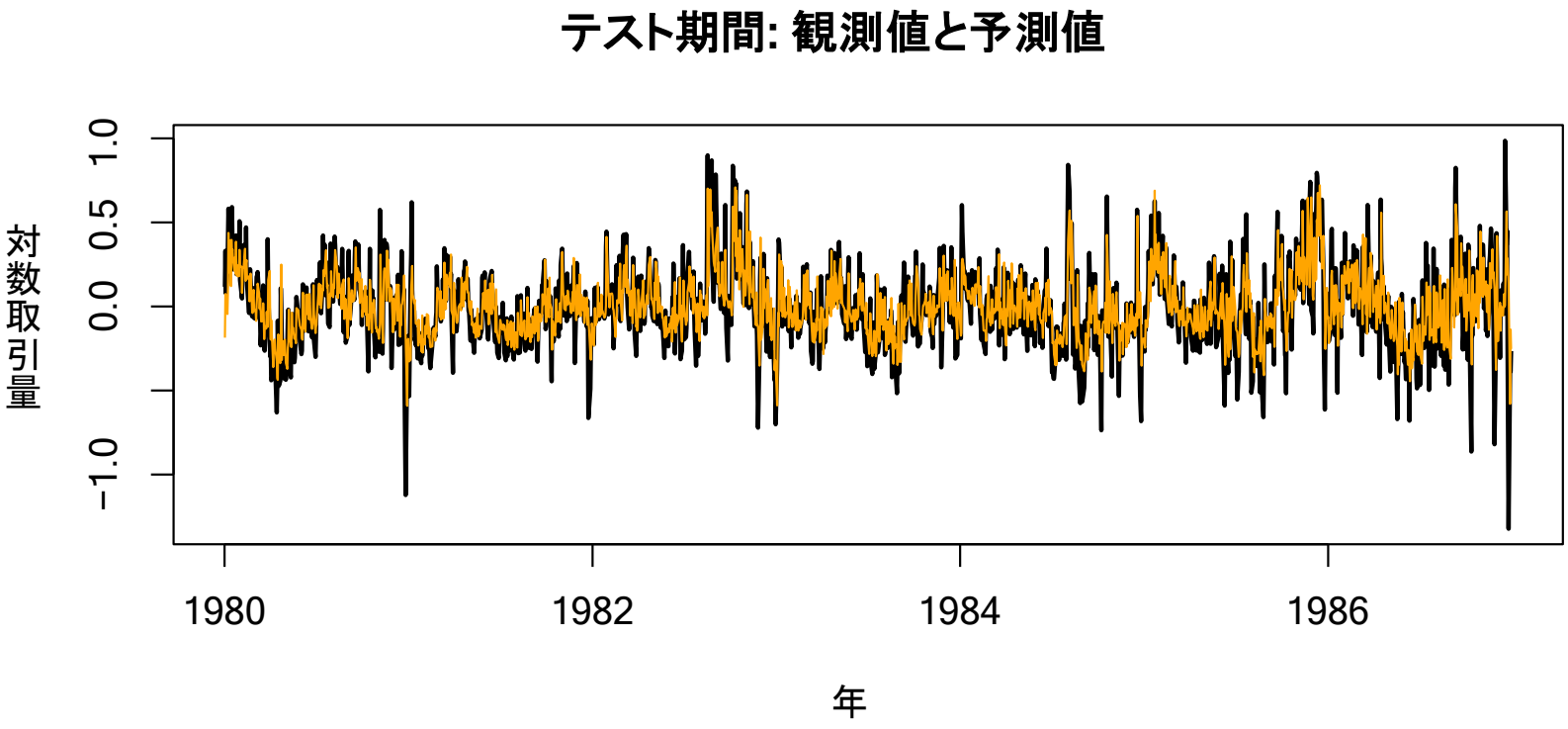
ラグと呼ばれるあらかじめ決めた長さLによって、$X = \{X_1, X_2, \ldots, X_L\}$という入力の短期の列を得る:

$$X_1 = \begin{pmatrix} v_{t-L} \\ r_{t-L} \\ z_{t-L} \end{pmatrix}, X_2 = \begin{pmatrix} v_{t-L+1} \\ r_{t-L+1} \\ z_{t-L+1} \end{pmatrix}, \ldots, X_L = \begin{pmatrix} v_{t-1} \\ r_{t-1} \\ z_{t-1} \end{pmatrix}, Y = v_t$$

$T = 6051$なので$L = 5$により6046のこのような(X, Y)のペアができる.

初めの4281個をトレーニングデータとして、続く1770個をテストデータとする. 12の隠れユニットを各ラグステップ毎(つまりA_ℓ毎)に持つRNNをフィットさせる

NYSEデータに対するRNNの結果



図はテスト期間の予測と真値を示している

RNNだと $R^2 = 0.42$
*straw man*だと $R^2 = 0.18$ —前日の対数取引量を
今日の予測値とする

自己回帰予測

RNNによる予測は伝統的な自己回帰による方法と似ている

$$\boldsymbol{y} = \begin{bmatrix} v_{L+1} \\ v_{L+2} \\ v_{L+3} \\ \vdots \\ v_T \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} 1 & v_L & v_{L-1} & \cdots & v_1 \\ 1 & v_{L+1} & v_L & \cdots & v_2 \\ 1 & v_{L+2} & v_{L+1} & \cdots & v_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & v_{T-1} & v_{T-2} & \cdots & v_{T-L} \end{bmatrix}$$

$\boldsymbol{y}$の$\boldsymbol{M}$への回帰によって、

$$\hat{v}_t = \hat{\beta}_0 + \hat{\beta}_1 v_{t-1} + \hat{\beta}_2 v_{t-2} + \cdots + \hat{\beta}_L v_{t-L}$$

オーダーLの自己回帰モデル、$AR(L)$として知られている.

NYSEデータに対しては、行列$\boldsymbol{M}$にダウジョーンズリターンと対数ボラティリティのラグバージョンを含めることができ、$3L+1$列の行列となる

NYSEデータに対する自己回帰の結果

$R^2 =$ 0.41 AR(5)モデル (16パラメータ)

$R^2 =$ 0.42 RNNモデル (205パラメータ)

$R^2 =$ 0.42 AR(5)モデルをニューラルネットワークによりフィット

$R^2 =$ 0.46 予測する曜日が含まれるとき、どんなモデルでも

RNNの要約

- RNNの最もシンプルなものを紹介した. 多くの複雑なバージョンがある
- 1つとしては、1次元の画像として列を扱い、CNNをフィットに用いる. 例えば、埋め込み表現を用いた単語の列は画像として捉え、列に沿って畳込みフィルタを適用できる
- 隠れ層を増やし、各隠れ層を列とし、前の隠れ層が入力となるように扱える
- 出力も列とすることができ、入出力が隠れユニットを共有するようにできる. seq2seq学習と呼ばれるものが言語学習に用いられる

いつ深層学習を用いるか

- CNNは画像分類やモデリングで大きな成功を収め、医療診断でも用いられ始めている. 例えば、デジタルマンモグラフィー、眼科、MRI検査、デジタルX線など

いつ深層学習を用いるか

- CNNは画像分類やモデリングで大きな成功を収め、医療診断でも用いられ始めている. 例えば、デジタルマンモグラフィー、眼科、MRI検査、デジタルX線など
- RNNは音声モデルや翻訳、予測に有用となっている

いつ深層学習を用いるか

- CNNは画像分類やモデリングで大きな成功を収め、医療診断でも用いられ始めている. 例えば、デジタルマンモグラフィー、眼科、MRI検査、デジタルX線など
- RNNは音声モデルや翻訳、予測に有用となっている

常に深層学習を用いるべきか？

いつ深層学習を用いるか

- CNNは画像分類やモデリングで大きな成功を収め、医療診断でも用いられ始めている. 例えば、デジタルマンモグラフィー、眼科、MRI検査、デジタルX線など
- RNNは音声モデルや翻訳、予測に有用となっている

常に深層学習を用いるべきか？

- 信号雑音比が大きい時、例えば画像認識や翻訳などでは大きく成功している. データセットが大きく、過適合が問題にならない

いつ深層学習を用いるか

- CNNは画像分類やモデリングで大きな成功を収め、医療診断でも用いられ始めている. 例えば、デジタルマンモグラフィー、眼科、MRI検査、デジタルX線など
- RNNは音声モデルや翻訳、予測に有用となっている

常に深層学習を用いるべきか？

- 信号雑音比が大きい時、例えば画像認識や翻訳などでは大きく成功している. データセットが大きく、過適合が問題にならない
- ノイズの大きなデータに対しては、よりシンプルなモデルが上手く行く
  - NYSEデータでは, AR(5)モデルはRNNよりシンプルで上手く働いている
  - IMDBレビューデータではglmnetによる線形モデルがニューラルネットと同程度、RNNよりは良い働きとなっている

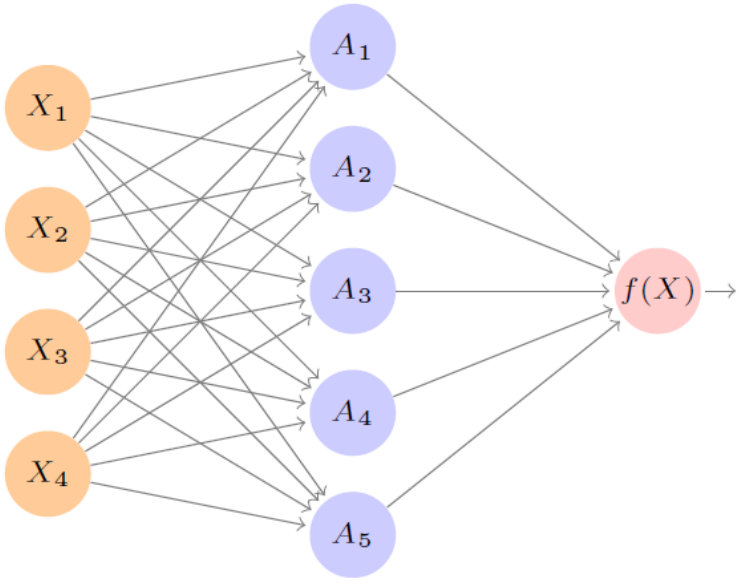
いつ深層学習を用いるか

- CNNは画像分類やモデリングで大きな成功を収め、医療診断でも用いられ始めている. 例えば、デジタルマンモグラフィー、眼科、MRI検査、デジタルX線など
- RNNは音声モデルや翻訳、予測に有用となっている

常に深層学習を用いるべきか？

- 信号雑音比が大きい時、例えば画像認識や翻訳などでは大きく成功している. データセットが大きく、過適合が問題にならない
- ノイズの大きなデータに対しては、よりシンプルなモデルが上手く行く
  - NYSEデータでは, AR(5)モデルはRNNよりシンプルで上手く働いている
  - IMDBレビューデータではglmnetによる線形モデルがニューラルネットと同程度、RNNよりは良い働きとなっている
- オッカムの剃刀の原則を支持する — 同程度に働くならば、よりシンプルな方が良い. より解釈がしやすくなる

ニューラルネットのフィット

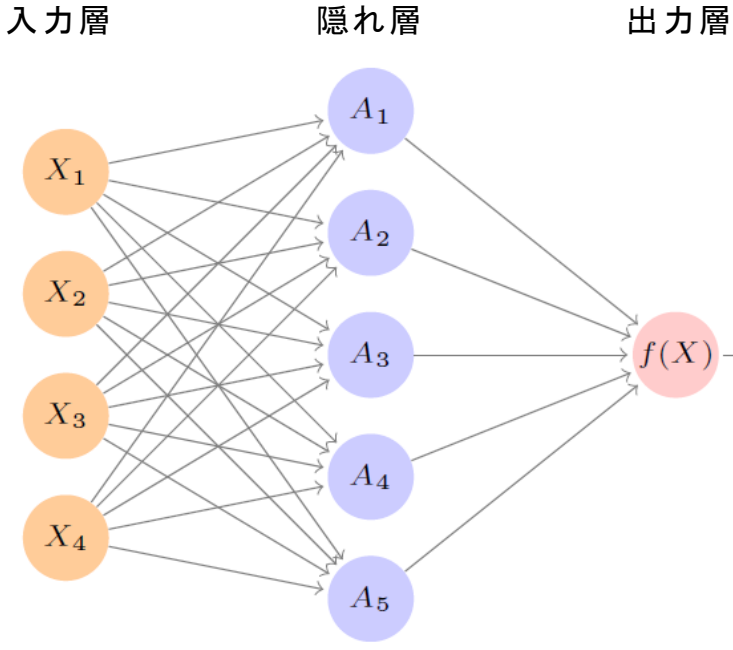


ただし

$$\underset{\{w_k\}_1^K,\beta}{\text{minimize}} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$$

$$f(x_i) = \beta_0 + \sum_{k=1}^{K}\beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij})$$

ニューラルネットのフィット



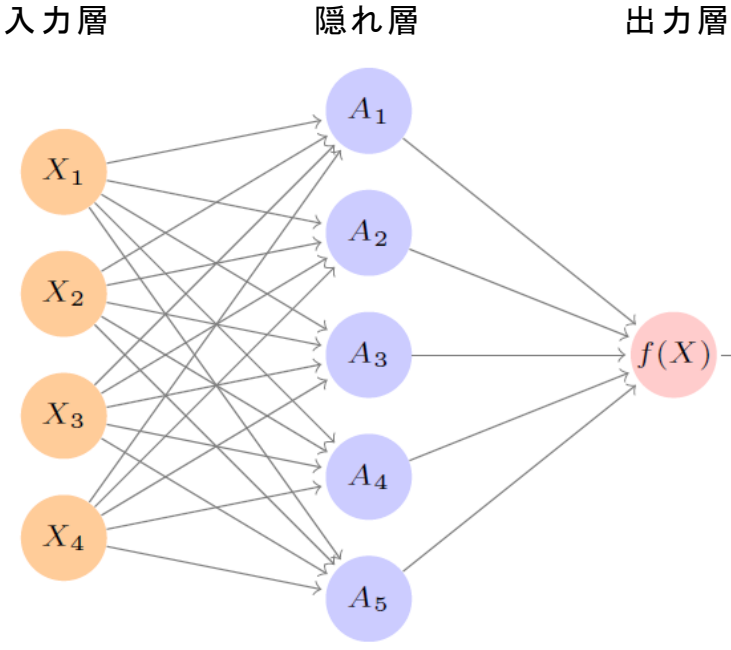
ただし

$$\underset{\{w_k\}_1^K,\beta}{\text{minimize}} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$$

$$f(x_i) = \beta_0 + \sum_{k=1}^{K}\beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij})$$

この問題は目的関数が非凸なので難しい

ニューラルネットのフィット



ただし

$$\underset{\{w_k\}_1^K,\beta}{\text{minimize}} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$$

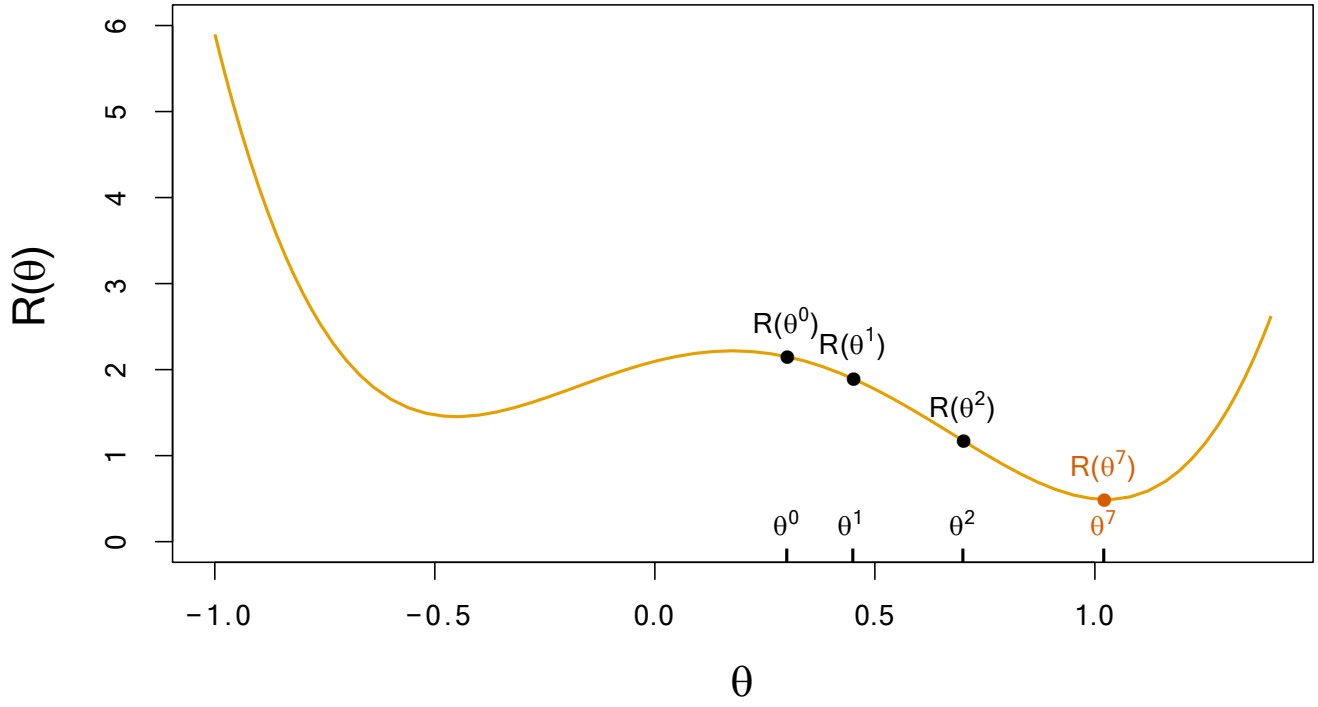
$$f(x_i) = \beta_0 + \sum_{k=1}^{K}\beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij})$$

この問題は目的関数が非凸なので難しい

にもかかわらず、効率的なアルゴリズムが開発され、
複雑なニューラルネットを効率的に最適化できる

非凸関数と勾配降下

$R(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - f_\theta(x_i)\right)^2, \theta = (\{w_k\}_1^K, \beta)$とする



1. θはθ^0で推測を始め、$t = 0$とする
2. 目的関数$R(\theta)$が減少する限り、以下を繰り返す:
   (a) θを少し変化させるδで$\theta^{t+1} = \theta^t + \delta$が目的関数を減少させるもの、つまり、$R(\theta^{t+1}) < R(\theta^t)$となるものを見つける
   (b) $t \leftarrow t + 1$とする

勾配降下(続き)

- 今回のシンプルな例では、大域的な最小値に達した
- θ^0の少し左から始めると逆の方向に進み、局所的な最小値で終了してしまう
- θが多次元だが、1次元で説明した. 次元が高い時、局所的最小値なのかどうかを確かめることも難しくなる

どうやって点を降下させる方向δを見つけるか. 勾配ベクトルを計算する

$$\nabla R(\theta^t) = \left.\frac{\partial R(\theta)}{\partial \theta}\right|_{\theta=\theta^t}$$

つまり、現在のθ^tでの偏微分のベクトル.

勾配は上昇方向を示すので、更新は$\delta = -\rho \nabla R(\theta^t)$で

$$\theta^{t+1} \leftarrow \theta^t - \rho \nabla R(\theta^t)$$

ただし、ρは学習率 (通常小さな値で、例えば$\rho = 0.001$)

勾配と逆誤差伝播

$R(\theta) = \sum_{i=1}^{n} R_i(\theta)$は和の形なので、勾配は勾配の和となる

$$R_i(\theta) = \frac{1}{2}\left(y_i - f_\theta(x_i)\right)^2 = \frac{1}{2}\left(y_i - \beta_0 - \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\right)\right)^2$$

記法の簡略化のため、$z_{ik} = w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}$

逆誤差伝播は微分の連鎖律を用いる

$$\begin{aligned}
\frac{\partial R_i(\theta)}{\partial \beta_k} &= \frac{\partial R_i(\theta)}{\partial f_\theta(x_i)} \cdot \frac{\partial f_\theta(x_i)}{\partial \beta_k} \\
&= -\left(y_i - f_\theta(x_i)\right) \cdot g(z_{ik}) \\
\frac{\partial R_i(\theta)}{\partial w_{kj}} &= \frac{\partial R_i(\theta)}{\partial f_\theta(x_i)} \cdot \frac{\partial f_\theta(x_i)}{\partial g(z_{ik})} \cdot \frac{\partial g(z_{ik})}{\partial z_{ik}} \cdot \frac{\partial z_{ik}}{\partial w_{kj}} \\
&= -\left(y_i - f_\theta(x_i)\right) \cdot \beta_k \cdot g'(z_{ik}) \cdot x_{ij}
\end{aligned}$$

コツ

- 遅い学習. 勾配降下がゆっくりで、学習率ρが小さければ更にゆっくりになる. 早期打ち切りとともに用いることで、これは正則化の1つになる.

コツ

- 遅い学習. 勾配降下がゆっくりで、学習率ρが小さければ更にゆっくりになる. 早期打ち切りとともに用いることで、これは正則化の1つになる.
- 確率勾配降下(SGD). 勾配を計算する際に、すべてのデータを使うよりも、各ステップでランダムに小さなミニバッチを用いる. 例えば、MNISTデータでは、$n = 6$万で128個のデータのミニバッチを用いる

コツ

- 遅い学習. 勾配降下がゆっくりで、学習率ρが小さければ更にゆっくりになる. 早期打ち切りとともに用いることで、これは正則化の1つになる.
- 確率勾配降下(SGD). 勾配を計算する際に、すべてのデータを使うよりも、各ステップでランダムに小さなミニバッチを用いる. 例えば、MNISTデータでは、$n = 6$万で128個のデータのミニバッチを用いる
- エポックは繰り返しの回数で、トータルでn個のサンプルが処理されるミニバッチの更新数に値する. つまり、MNISTでは$6万/128 \approx 469$

コツ

- 遅い学習. 勾配降下がゆっくりで、学習率ρが小さければ更にゆっくりになる. 早期打ち切りとともに用いることで、これは正則化の1つになる.
- 確率勾配降下(SGD). 勾配を計算する際に、すべてのデータを使うよりも、各ステップでランダムに小さなミニバッチを用いる. 例えば、MNISTデータでは、$n = 6$万で128個のデータのミニバッチを用いる
- エポックは繰り返しの回数で、トータルでn個のサンプルが処理されるミニバッチの更新数に値する. つまり、MNISTでは $6\text{万}/128 \approx 469$
- 正則化. リッジやLasso正則化は各層で重みの縮小のために用いられる. 正則化の他の方法として2つ挙げると、ドロップアウトや拡張がある. 次に議論する

ドロップアウト学習

- 各SGDの更新では、確率ϕでユニットを除き、$1/(1-\phi)$を重みにかけ、スケールを調整する

ドロップアウト学習

- 各SGDの更新では、確率ϕでユニットを除き、$1/(1-\phi)$を重みにかけ、スケールを調整する
- 単純な線形回帰では、この過程はリッジ正則化と同等

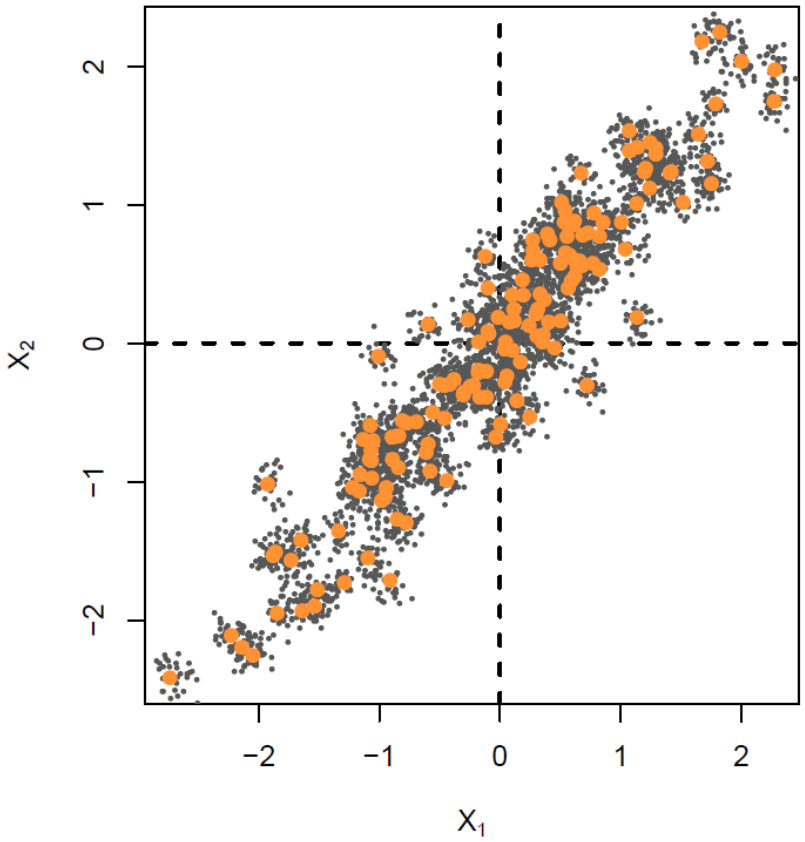
ドロップアウト学習

- 各SGDの更新では、確率ϕでユニットを除き、$1/(1-\phi)$を重みにかけ、スケールを調整する
- 単純な線形回帰では、この過程はリッジ正則化と同等
- リッジのように、他のユニットが一時的に除かれたものを代表し、それらの重みは近くのものによる

ドロップアウト学習

- 各SGDの更新では、確率ϕでユニットを除き、$1/(1-\phi)$を重みにかけ、スケールを調整する
- 単純な線形回帰では、この過程はリッジ正則化と同等
- リッジのように、他のユニットが一時的に除かれたものを代表し、それらの重みは近くのものによる
- ランダムフォレストで木を増やす場合にランダムに変数を除去することに似ている(8章)

リッジとデータ拡張



- (x_i, y_i)の多くのコピーを作る. x_iに対して小さなガウスノイズを加える. y_iのコピーはそのままにする
- x_iの小さな摂動にロバストにフィットさせることになり、最小二乗の場合ではリッジ正則化と同等

簡単なデータ拡張

- データ拡張は特にSGDで有効である. ここではCNNと画像分類に関して説明する

簡単なデータ拡張

- データ拡張は特にSGDで有効である. ここではCNNと画像分類に関して説明する
- SGDで抽出された時、各訓練画像は自然な変換がされる. 最終的には各もとの画像の周りに画像ができる

簡単なデータ拡張

- データ拡張は特にSGDで有効である. ここではCNNと画像分類に関して説明する
- SGDで抽出された時、各訓練画像は自然な変換がされる. 最終的には各もとの画像の周りに画像ができる
- ラベルはそのままで、tigerのまま

簡単なデータ拡張

- データ拡張は特にSGDで有効である. ここではCNNと画像分類に関して説明する
- SGDで抽出された時、各訓練画像は自然な変換がされる. 最終的には各もとの画像の周りに画像ができる
- ラベルはそのままで、tigerのまま
- CNNのパフォーマンスを改善し、リッジと同程度

二重降下

- ニューラルネットワークでは、隠れユニットが少な過ぎるよりも多過ぎる方が良いようである

二重降下

- ニューラルネットワークでは、隠れユニットが少な過ぎるよりも多過ぎる方が良いようである
- 同様に隠れ層も少ないより多い方が良い

二重降下

- ニューラルネットワークでは、隠れユニットが少な過ぎるよりも多過ぎる方が良いようである
- 同様に隠れ層も少ないより多い方が良い
- 確率勾配降下によって訓練誤差を0に近づける事が、訓練サンプル以外での誤差も良くすることが多い

二重降下

- ニューラルネットワークでは、隠れユニットが少な過ぎるよりも多過ぎる方が良いようである
- 同様に隠れ層も少ないより多い方が良い
- 確率勾配降下によって訓練誤差を0に近づける事が、訓練サンプル以外での誤差も良くすることが多い
- ユニットや層の数を増やし訓練誤差を0に近づけることで、訓練サンプル以外での誤差をより良くする

二重降下

- ニューラルネットワークでは、隠れユニットが少な過ぎるよりも多過ぎる方が良いようである
- 同様に隠れ層も少ないより多い方が良い
- 確率勾配降下によって訓練誤差を0に近づける事が、訓練サンプル以外での誤差も良くすることが多い
- ユニットや層の数を増やし訓練誤差を0に近づけることで、訓練サンプル以外での誤差をより良くする

過適合や通常のバイアス-バリアンスのトレードオフはどうなっているのか？

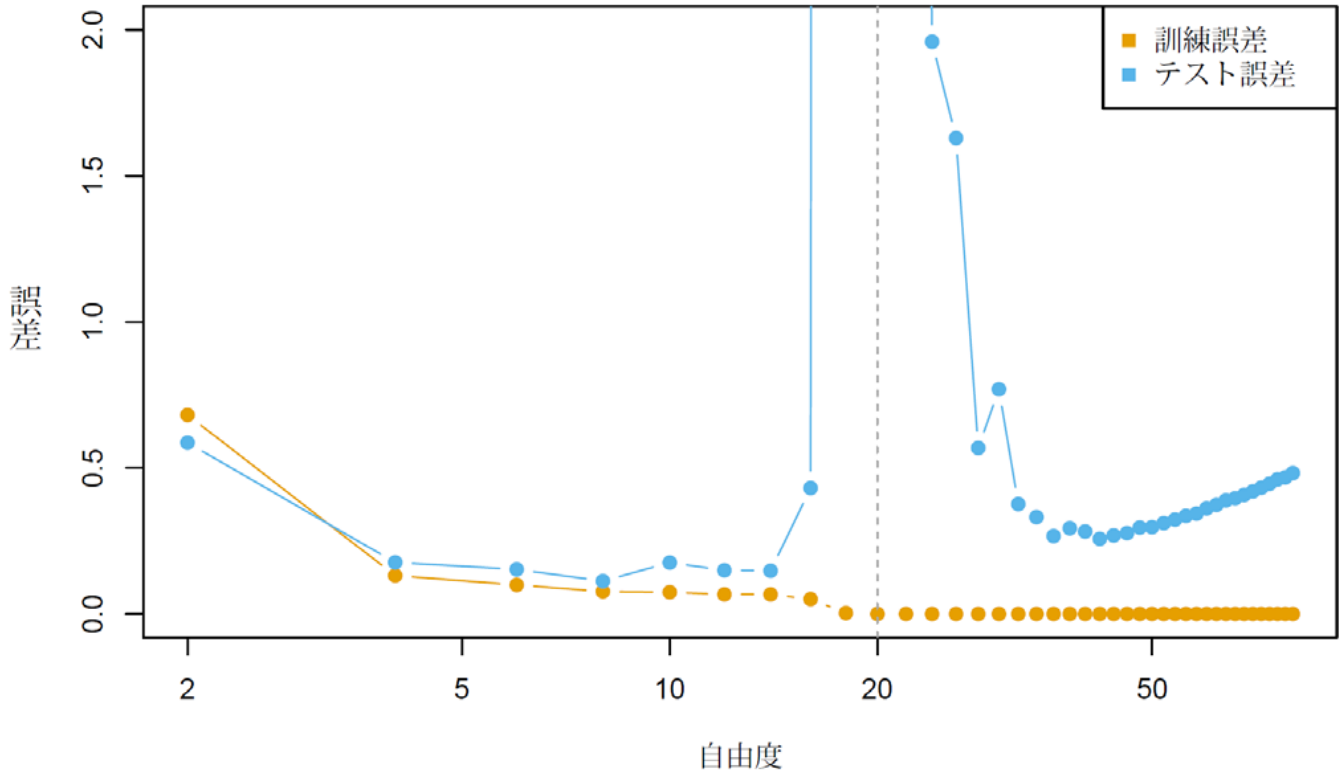
Belkin, Hsu, Ma and Mandal (arXiv 2018) *Reconciling Modern Machine Learning and the Bias-Variance Trade-off.*

(Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32):15849–15854, 2019.)

シミュレーション

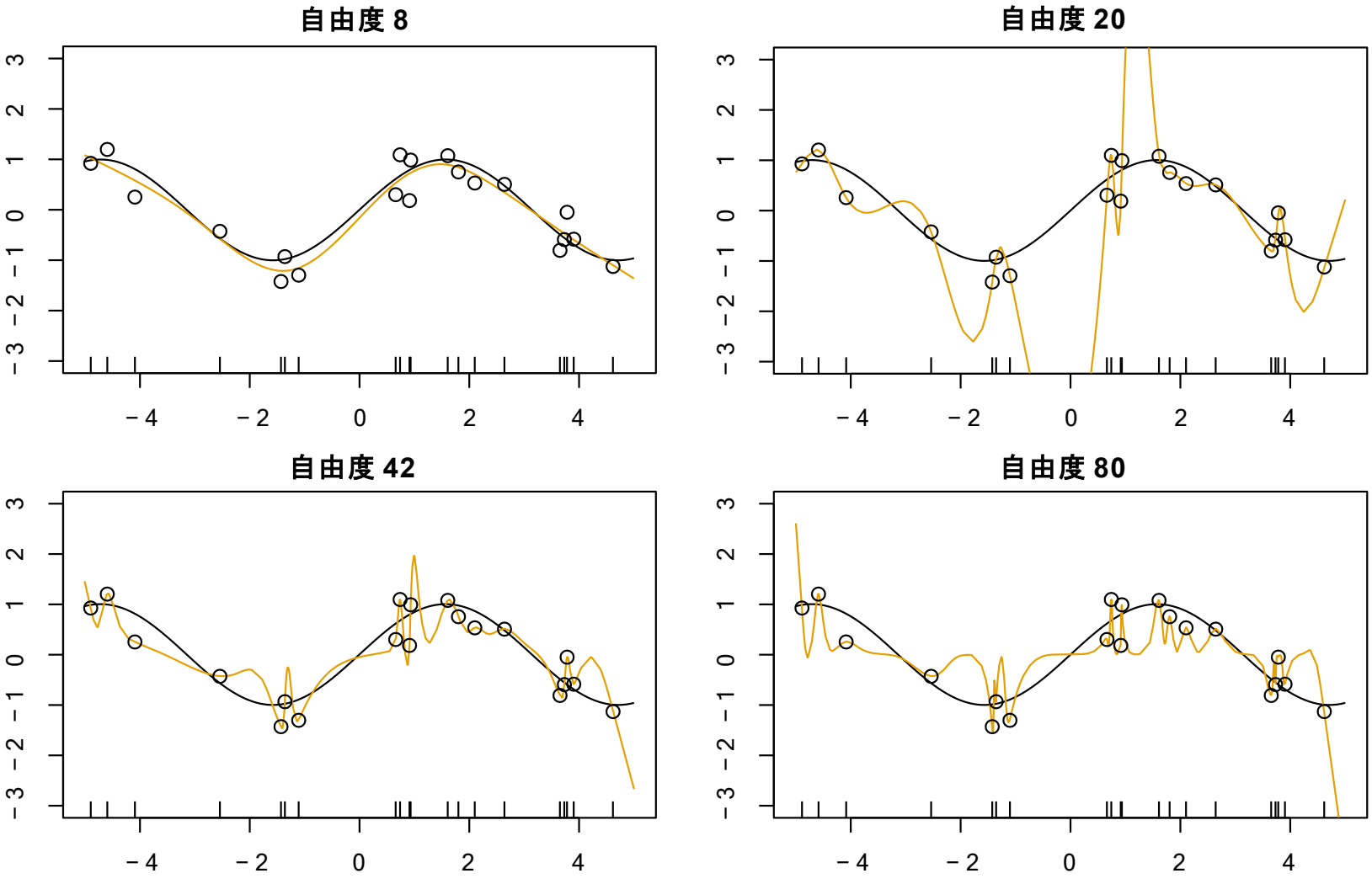
- $y = \sin(x) + \epsilon, x \sim U[-5,5]$で$\epsilon$は標準偏差が$0.3$のガウス
- 訓練セットは$n = 20$、テストセットは1万
- データに自由度dの自然スプラインをフィットさせる. つまり、d個の基底関数への線形回帰 : $\hat{y}_i = \hat{\beta}_1 N_1(x_i) + \hat{\beta}_2 N_2(x_i) + \cdots + \hat{\beta}_d N_d(x_i)$
- $d = 20$のとき、データを正確に学習し残差は0となる
- $d > 20$のとき、データに正確にフィットするがその解は一意ではない. 残差を0にする解の中でノルムを最小にする、つまり$\sum_{j=1}^{d} \hat{\beta}_j^2$を最小にするものを選ぶ

二重降下損失曲線



- $d \leq 20$のとき、最小二乗で、通常バイアス-バリアンストレードオフが見られる
- $d > 20$のとき、ノルム最小化に至る. dが20より大きくなる程、損失を0にすることは容易になるので、$\sum_{j=1}^{d} \hat{\beta}_j^2$は減少する. 結果としてより曲がらない解を得る

より曲がらない解



$d = 20$で、データを捉えた残差が0となる解を得る. dが大きい程簡単になる

いくつか知られたこと

- 大きな線形モデル($p \gg n$)で最小二乗によりフィットさせると、小さなステップサイズでのSGDはノルム最小化で残差を0にする解を導く

いくつか知られたこと

- 大きな線形モデル($p \gg n$)で最小二乗によりフィットさせると、小さなステップサイズでのSGDはノルム最小化で残差を0にする解を導く
- 確率的勾配*フロー* — つまりSGDの解のパス— はリッジのパスに似ていることがある

いくつか知られたこと

- 大きな線形モデル($p \gg n$)で最小二乗によりフィットさせると、小さなステップサイズでのSGDはノルム最小化で残差を0にする解を導く
- 確率的勾配*フロー* — つまりSGDの解のパス— はリッジのパスに似ていることがある
- アナロジーにより、深くて大きなニューラルネットをSGDによりフィットさせ訓練誤差を0に近づけると汎化の良い解を与える

いくつか知られたこと

- 大きな線形モデル($p \gg n$)で最小二乗によりフィットさせると、小さなステップサイズでのSGDはノルム最小化で残差を0にする解を導く
- 確率的勾配*フロー* — つまりSGDの解のパス— はリッジのパスに似ていることがある
- アナロジーにより、深くて大きなニューラルネットをSGDによりフィットさせ訓練誤差を0に近づけると汎化の良い解を与える
- 特に信号雑音比が高いとき — 例えば、画像認識 — 過適合になりにくく、誤差が0の解はほとんど信号、シグナルである

ソフトウェア

- ニューラルネットワークや深層学習には、素晴らしいソフトウェアを用いる事ができる. GoogleによるTensorflowやFacebookによるPyTorchがある. ともにPythonのパッケージ
- 10章のlabでは、Rで tensorflowとkerasパッケージを用いた実行例がある. ともにPythonに通じるようなものである. 教科書やオンラインの資料を見ると、これらの例や2版の教科書のlabの、RmarkdownやJupyter ノートブックがある
- Rのtorchパッケージも PyTorchを実行できる. 10章のlabはこれも利用できる. 資料ページとしては次を参照
  www.statlearning.com

第11章:生存時間解析と打ち切り
-Survival Analysis and Censoring-

- 打ち切り時刻-censoring-
- 生存関数(survival function),生存曲線(curve)
- カプラン・マイヤー(Kaplan-Meier)推定
- ログランク検定
- ハザード関数-Hazard Function-
- 比例ハザード(proportional hazard)モデル
- 部分尤度-partial likelihood-
- 修正生存曲線-Adjusted Survival Curves-

生存時間解析

- 生存時間解析では特別な目的変数: イベントが生きるまでの時間(time until an event occurs)を扱う.

生存時間解析

- 生存時間解析では特別な目的変数: イベントが生きるまでの時間 *(time until an event occurs)*を扱う.
- 例えば癌治療の患者について5年間の医療研究を行っていることを想定しよう.

生存時間解析

- 生存時間解析では特別な目的変数: イベントが生きるまでの時間 *(time until an event occurs)*を扱う.
- 例えば癌治療の患者について5年間の医療研究を行っていることを想定しよう.
- 患者についてベースラインの健康測定値, 治療のタイプなどの要素を使い, 患者の生存時間を予測するためのモデルをフィットしたいとする.

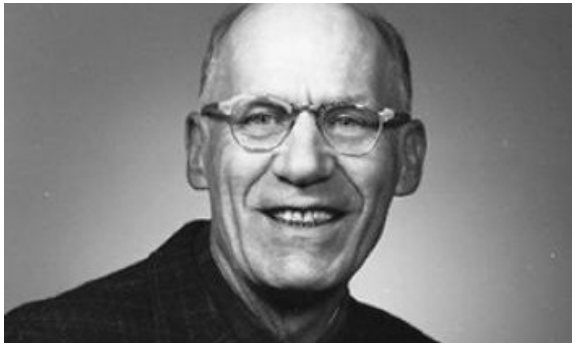
生存時間解析

- 生存時間解析では特別な目的変数: イベントが生きるまでの時間*(time until an event occurs)*を扱う.
- 例えば癌治療の患者について5年間の医療研究を行っていることを想定しよう.
- 患者についてベースラインの健康測定値, 治療のタイプなどの要素を使い, 患者の生存時間を予測するためのモデルをフィットしたいとする.
- 回帰問題*(regression problem)*のように聞こえるが, 一つの重要な複雑性が存在する: 何人かの患者は研究の最後まで生き延びることである. そうした患者の生存時間は打ち切り*(censored)*されているという.

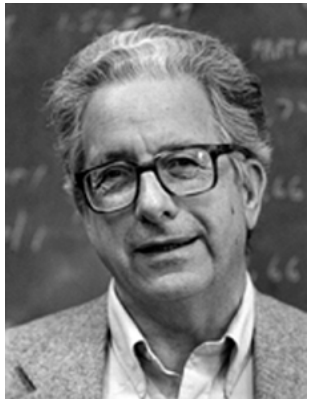
生存時間解析

- 生存時間解析では特別な目的変数: イベントが生きるまでの時間*(time until an event occurs)*を扱う.
- 例えば癌治療の患者について5年間の医療研究を行っていることを想定しよう.
- 患者についてベースラインの健康測定値, 治療のタイプなどの要素を使い, 患者の生存時間を予測するためのモデルをフィットしたいとする.
- 回帰問題*(regression problem)*のように聞こえるが, 一つの重要な複雑性が存在する: 何人かの患者は研究の最後まで生き延びることである. そうした患者の生存時間は打ち切り*(censored)*されているという.
- こうした生き残った患者集団データを捨てるべきではない. と云うのは少なくとも5年間生き残れたというというのは貴重な情報なのである.

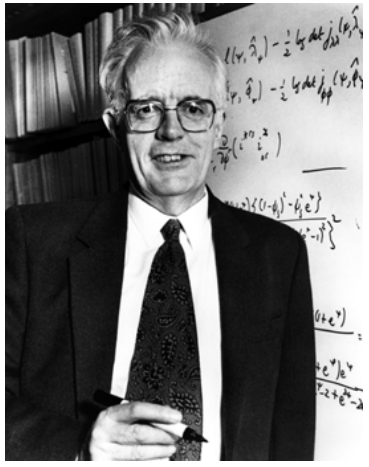
生存時間解析の主要な貢献者



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

William Haenszel

(***log rank test***)

Terry Therneau

(author of Survival package in R)

非医療分野の例

- 生存時間解析の応用は医療分野に限らない. 例えば, ある企業が客がサービスの契約をキャンセルする事象：解約(*churn*)をモデル分析ことを考えよう.

非医療分野の例

- 生存時間解析の応用は医療分野に限らない. 例えば, ある企業が客がサービスの契約をキャンセルする事象：解約*(churn)*をモデル分析ことを考えよう.
- 企業はある機関に顧客のデータを集め, 顧客がキャンセルする時刻を予測しようとするかもしれない.

非医療分野の例

- 生存時間解析の応用は医療分野に限らない. 例えば, ある企業が客がサービスの契約をキャンセルする事象：解約(*churn*)をモデル分析ことを考えよう.
- 企業はある機関に顧客のデータを集め, 顧客がキャンセルする時刻を予測しようとするかもしれない.
- しかしながら, データを集めた終了時刻までにすべての顧客が購買契約をキャンセルするわけではない, そうした顧客キャンセルする時刻は打ち切られていると見なせる.

非医療分野の例

- 生存時間解析の応用は医療分野に限らない. 例えば, ある企業が客がサービスの契約をキャンセルする事象：解約(*churn*)をモデル分析ことを考えよう.
- 企業はある機関に顧客のデータを集め, 顧客がキャンセルする時刻を予測しようとするかもしれない.
- しかしながら, データを集めた終了時刻までにすべての顧客が購買契約をキャンセルするわけではない, そうした顧客キャンセルする時刻は打ち切られていると見なせる.
- 生存時間解析は統計学の中でも非常によく研究されている分野である. しかしながら機械学習machine learningの研究者の間ではあまり注目されてはいない分野となっている.

生存時間と打ち切り時刻
Survival and Censoring Times

- 各個体について, 真の失敗生起時刻*(failure)*, あるいは事象*(event)*の時刻T, および真の打ち切り時刻Cが存在するとしよう.

生存時間と打ち切り時刻
Survival and Censoring Times

- 各個体について, 真の失敗生起時刻*(failure)*, あるいは事象*(event)*の時刻T, および真の打ち切り時刻Cが存在するとしよう.
- 生存時間(survival time)は関心のある事象が起きる(例えば死亡など)が起きる時刻を表そう.

生存時間と打ち切り時刻
Survival and Censoring Times

- 各個体について, 真の失敗生起時刻*(failure)*, あるいは事象*(event)*の時刻T, および真の打ち切り時刻Cが存在するとしよう.
- 生存時間(survival time)は関心のある事象が起きる(例えば死亡など)が起きる時刻を表そう.
- 次に打ち切り時刻*(censoring)*とは打ち切りが生じた時刻; 例えば患者が途中にいなくなった時刻, あるいは研究終了時である.

生存時間と打ち切り– 続き

- 観察するのは生存時刻(survival time) T, あるいは打ち切り時刻(censoring time) Cとする. 特に次の確率変数の実現値を観測

$$Y = \min(T, C)$$

生存時間と打ち切り– 続き

- 観察するのは生存時刻(survival time) T, あるいは打ち切り時刻(censoring time) Cとする. 特に次の確率変数の実現値を観測

$$Y = \min(T, C)$$

- 事象(event)が打ち切り時刻の前に起きる (すなわち$T < C$)ならば, 真の生存時刻Tを観測できる; 事象が起きる前に打ち切りが起きる ($T > C$)ならば打ち切り時刻が観測される. 状態の指示関数を観測すると見なせるが,

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C \end{cases}$$

生存時間と打ち切り– 続き

- 観察するのは生存時刻(survival time) T, あるいは打ち切り時刻(censoring time) Cとする. 特に次の確率変数の実現値を観測

$$Y = \min(T, C)$$

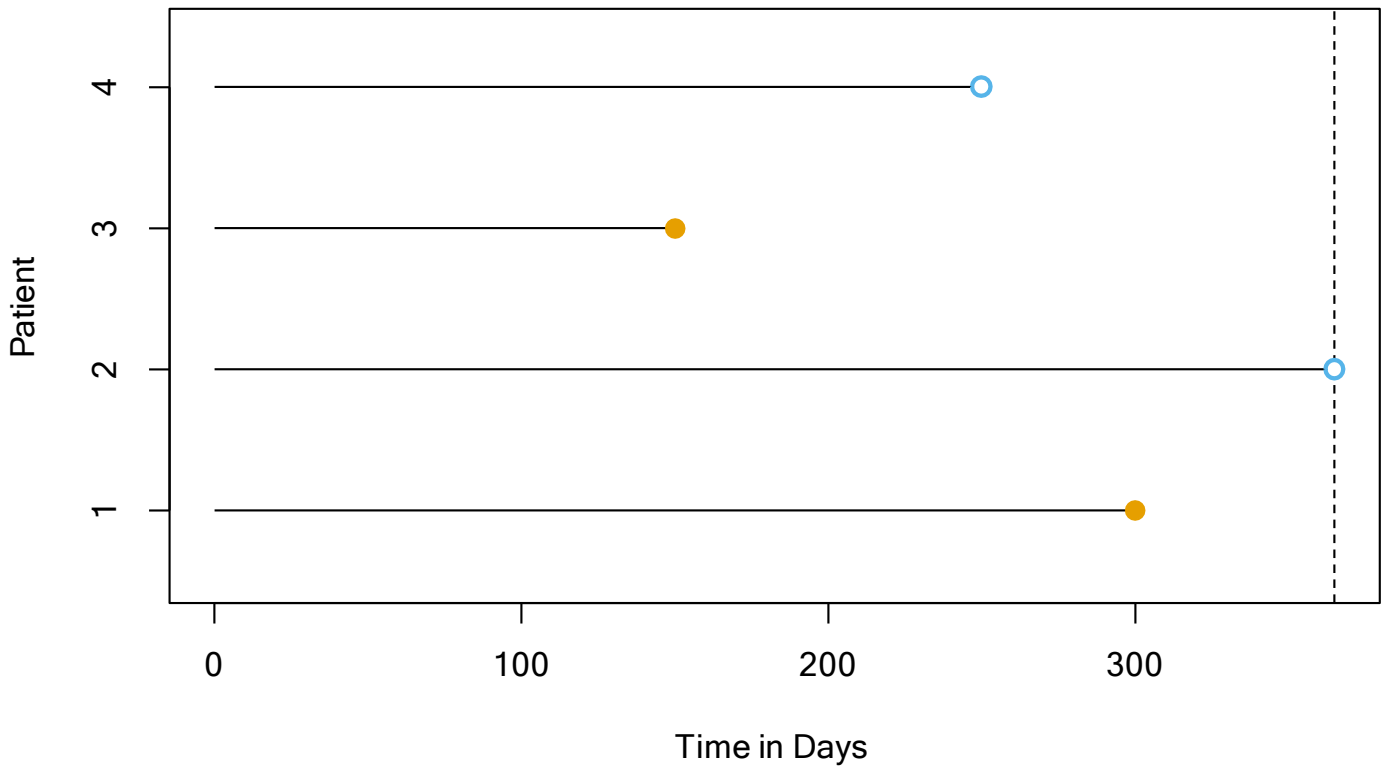
- 事象(event)が打ち切り時刻の前に起きる (すなわち$T < C$)ならば, 真の生存時刻Tを観測できる; 事象が起きる前に打ち切りが起きる ($T > C$)ならば打ち切り時刻が観測される. 状態の指示関数を観測すると見なせるが,

$$\delta = \begin{cases} 1 & \text{if } T \leq C \\ 0 & \text{if } T > C \end{cases}$$

- したがって, データとして観察するのはn個の組 $(Y, \delta), (y_1, \delta_1), \ldots, (y_n, \delta_n)$となる.

例示

打ち切りがある生存時間データを例示する. 患者1と患者3は事象を観測. 患者2は研究の終了時にも生存, 患者4は途中でドロップアウトしている.



打ち切りの考察

- ガン研究に参加したかなりの患者が病状が悪化したのでドロップアウトしたとしよう.

打ち切りの考察

- 癌研究に参加したかなりの患者が病状が悪化したのでドロップアウトしたとしよう.
- このとき患者がなぜドロップアウトしたかの理由を考慮することがない分析では真の平均生存時間を過大推定しがちであろう.

打ち切りの考察

- 癌研究に参加したかなりの患者が病状が悪化したのでドロップアウトしたとしよう.
- このとき患者がなぜドロップアウトしたかの理由を考慮することがない分析では真の平均生存時間を過大推定しがちであろう.
- 同様に, 病状が悪化した男性患者が病状の悪化した女性患者よりも研究から脱落しがちであったとしよう. このとき男性と女性の生存曲線を比較すると 男性の生存時間は女性の生存時間よりも長いことを誤って示唆することになる.

打ち切りの考察

- 癌研究に参加したかなりの患者が病状が悪化したのでドロップアウトしたとしよう.
- このとき患者がなぜドロップアウトしたかの理由を考慮することがない分析では真の平均生存時間を過大推定しがちであろう.
- 同様に, 病状が悪化した男性患者が病状の悪化した女性患者よりも研究から脱落しがちであったとしよう. このとき男性と女性の生存曲線を比較すると 男性の生存時間は女性の生存時間よりも長いことを誤って示唆することになる.
- 一般には特徴(features)に条件付けして, 事象の発生時刻Tは打ち切り時刻Cと独立性*(independent)*を仮定する必要がある. しかしここで述べた二つの例では独立な打ち切りという仮定が成り立っていない.

生存曲線
-Survival Curve-

- 生存関数(survival function),生存曲線(curve) は次のように定義される

$$S(t) = \Pr(T > t).$$

生存曲線
-Survival Curve-

- 生存関数(survival function),生存曲線(curve) は次のように定義される

$$S(t) = \Pr(T > t).$$

- 減少関数により過去の時刻tに生存している確率を表現している.

生存曲線
-Survival Curve-

- 生存関数(survival function),生存曲線(curve) は次のように定義される

$$S(t) = \Pr(T > t).$$

- 減少関数により過去の時刻tに生存している確率を表現している.
- 例えばある企業が顧客のキャンセル行動をモデル化したいとしよう. 時刻Tにより顧客が購買をキャンセルする時刻とする.

生存曲線
-Survival Curve-

- 生存関数(survival function),生存曲線(curve) は次のように定義される

$$S(t) = \Pr(T > t).$$

- 減少関数により過去の時刻tに生存している確率を表現している.
- 例えばある企業が顧客のキャンセル行動をモデル化したいとしよう. 時刻Tにより顧客が購買をキャンセルする時刻とする.
- このとき$S(t)$はある顧客が時刻tより先にキャンセルする確率を表す. $S(t)$が大きければある顧客が時刻tより前にキャンセルする可能性は低い.

生存曲線の推定
Estimating the Survival Curve

- 脳腫瘍BrainCancerデータを考える.原発性脳腫瘍(primary brain tumors)に罹患,定位放射線治療(stereotactic radiation methods)を行っているで患者の生存時間としよう.

(訳注)医学専門用語については訳者など非専門家には正確な理解は困難である。

生存曲線の推定
Estimating the Survival Curve

- 脳腫瘍BrainCancerデータを考える.原発性脳腫瘍(primary brain tumors)に罹患,定位放射線治療(stereotactic radiation methods)を行っているで患者の生存時間としよう.
- 予測変数はgtv(肉眼的腫瘍体積, gross tumor volume,立方センチメートル); sex(性別m 男性or女性); diagnosis(meningioma, LG glioma, HG glioma, その他); loc(腫瘍位置;テント下腫瘍or上部infratentorial,supratentorial); ki(カルノフスキー指数Karnofsky index); stereo(方位固定法stereotactic method).

生存曲線の推定
Estimating the Survival Curve

- 脳腫瘍BrainCancerデータを考える.原発性脳腫瘍(primary brain tumors)に罹患,定位放射線治療(stereotactic radiation methods)を行っているで患者の生存時間としよう.
- 予測変数はgtv(肉眼的腫瘍体積, gross tumor volume,立方センチメートル); sex(性別m 男性or女性); diagnosis(meningioma, LG glioma, HG glioma, その他); loc(腫瘍位置;テント下腫瘍or上部infratentorial,supratentorial); ki(カルノフスキー指数Karnofsky index); stereo(方位固定法stereotactic method).
- 88名の患者の内研究終了時に53名が生存.

生存曲線の推定 – 続き

- ここで次の量を推定したいとしよう$S(20) = \Pr(T > 20)$, 患者が少なくとも20か月は生きている確率とする.

生存曲線の推定 – 続き

- ここで次の量を推定したいとしよう$S(20) = \Pr(T > 20)$, 患者が少なくとも20か月は生きている確率とする.
- ここで単純に過去20ヶ月生存していた患者の割合を計算しがちである, つまり$Y > 20$である患者の割合である.

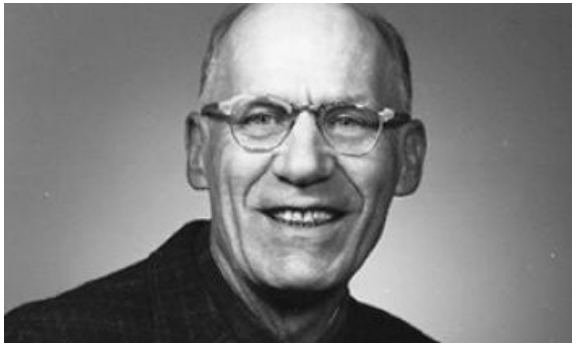
生存曲線の推定 – 続き

- ここで次の量を推定したいとしよう$S(20) = \Pr(T > 20)$, 患者が少なくとも20か月は生きている確率とする.
- ここで単純に過去20ヶ月生存していた患者の割合を計算しがちである, つまり$Y > 20$である患者の割合である.
- この数値は48/88となるので近似的に55% である.

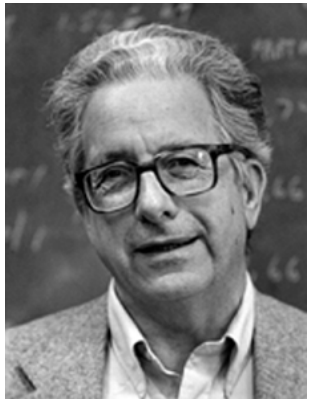
生存曲線の推定 – 続き

- ここで次の量を推定したいとしよう$S(20) = \Pr(T > 20)$, 患者が少なくとも20か月は生きている確率とする.
- ここで単純に過去20ヶ月生存していた患者の割合を計算しがちである, つまり$Y > 20$である患者の割合である.
- この数値は48/88となるので近似的に55% である.
- しかしながら, これは正しくないようである: 40名中の17名の患者が生存していないが, 実際には打ち切られている, この分析は20月以前に死亡していると仮定していることになる. したがって, 確率を過小推定している.

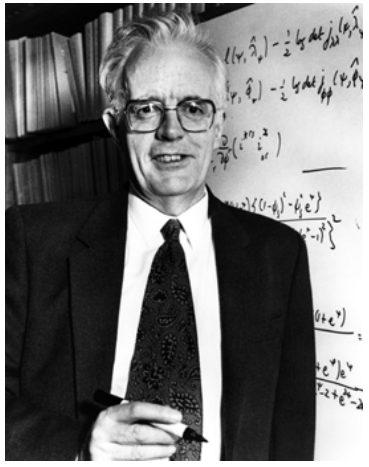
生存時間解析の主要な貢献者



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

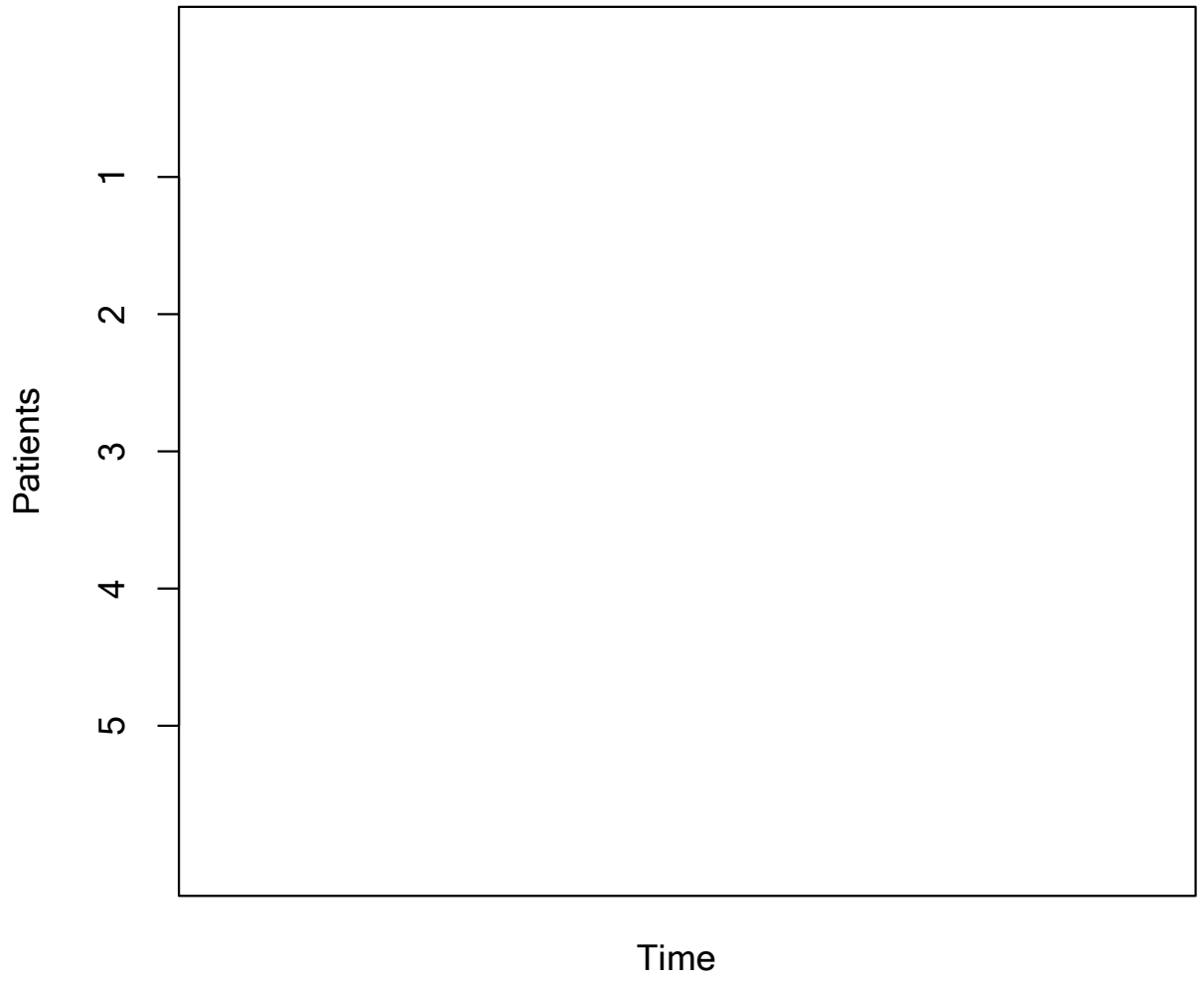
William Haenszel

(***log rank test***)

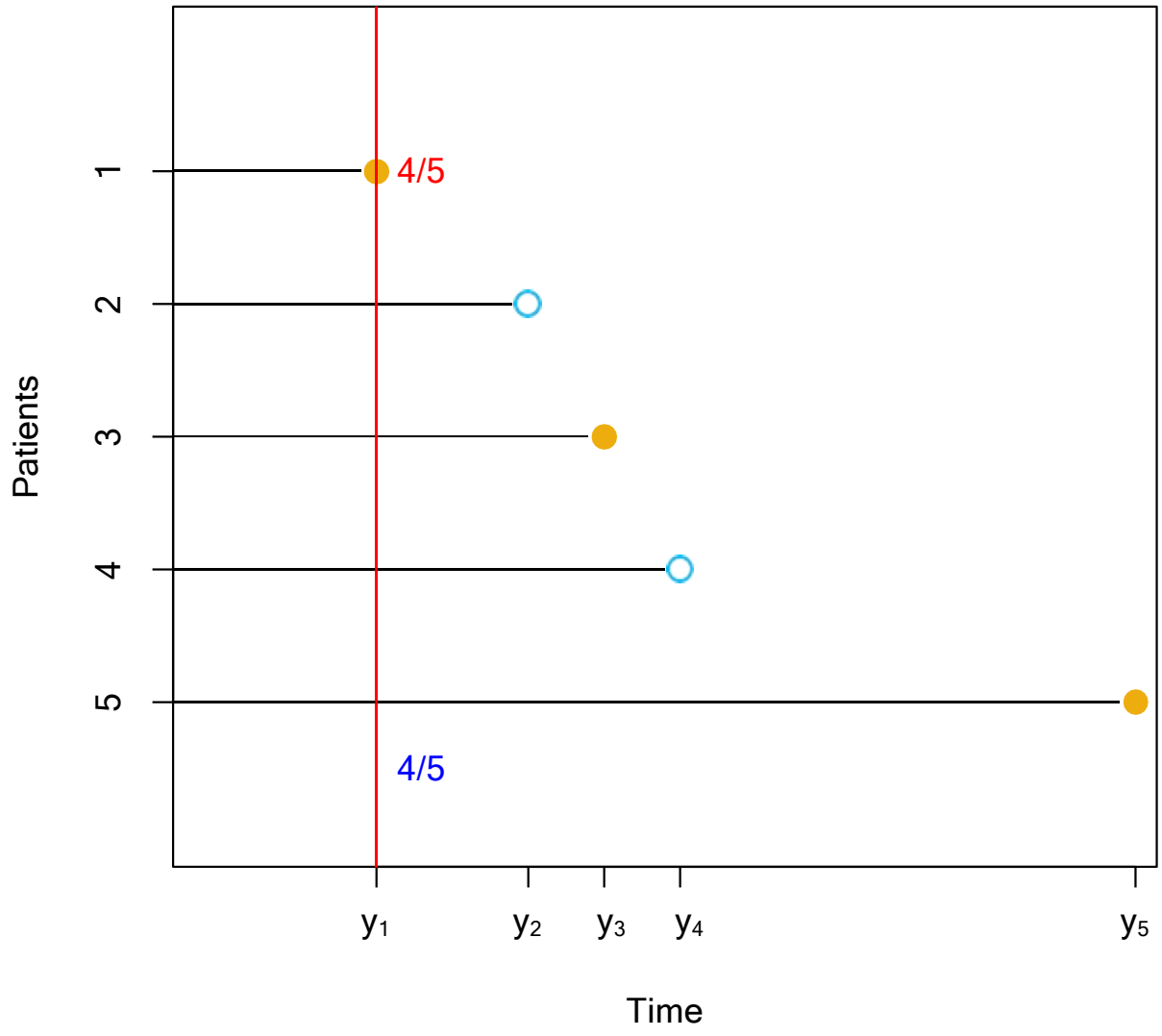
Terry Therneau

(author of Survival package in R)

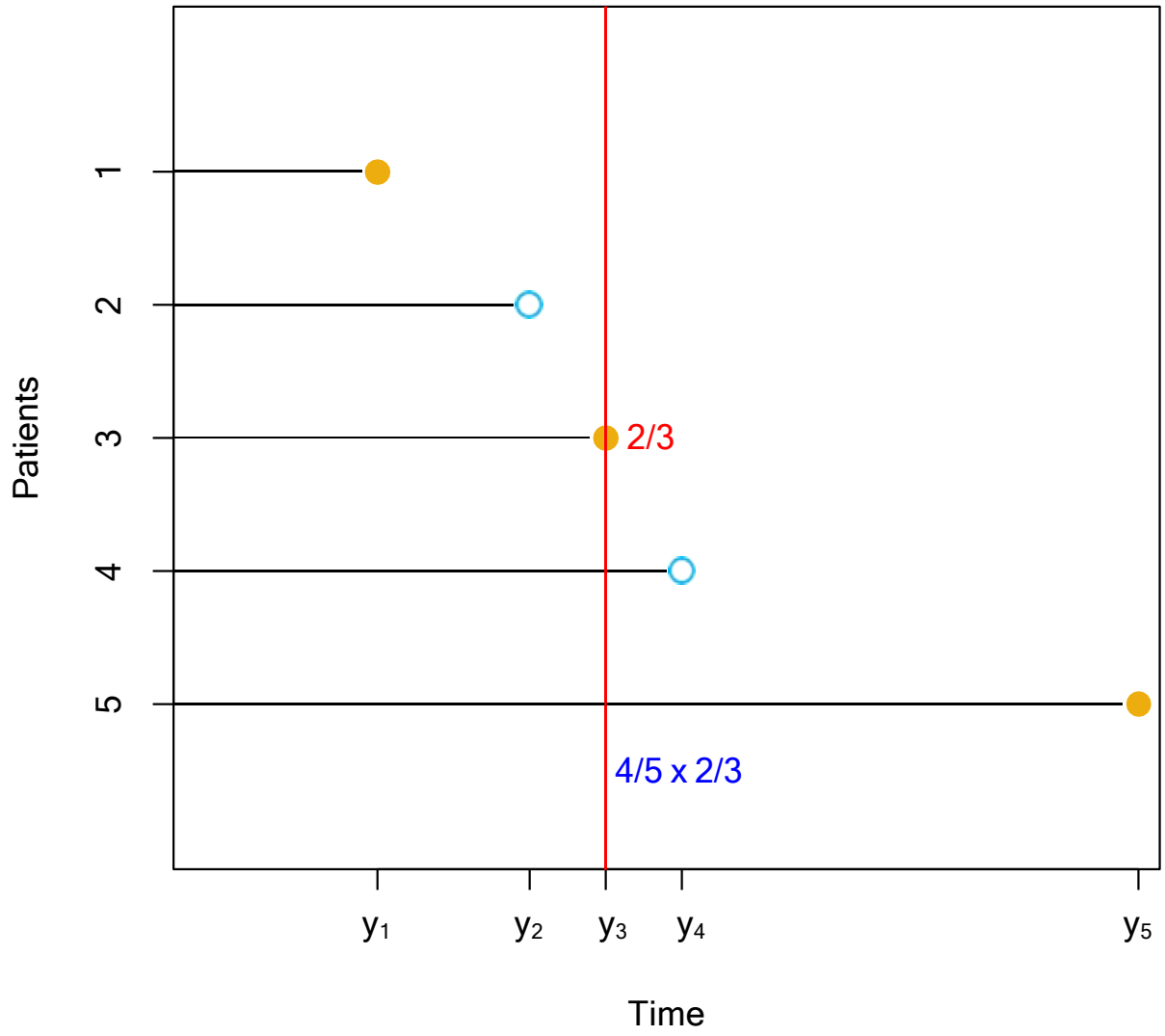
カプラン・マイヤー(Kaplan-Meier)推定



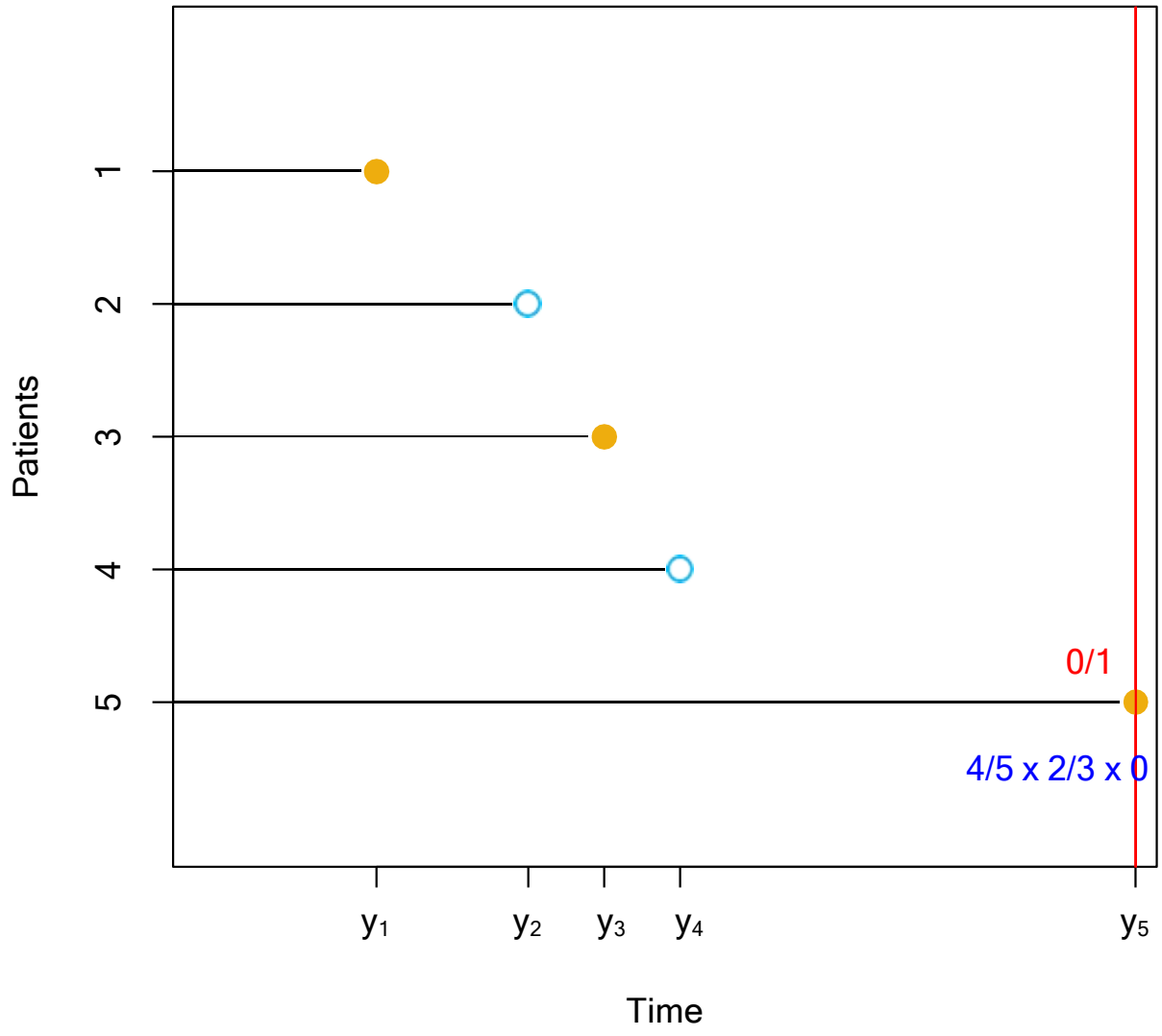
最初の失敗生起事象



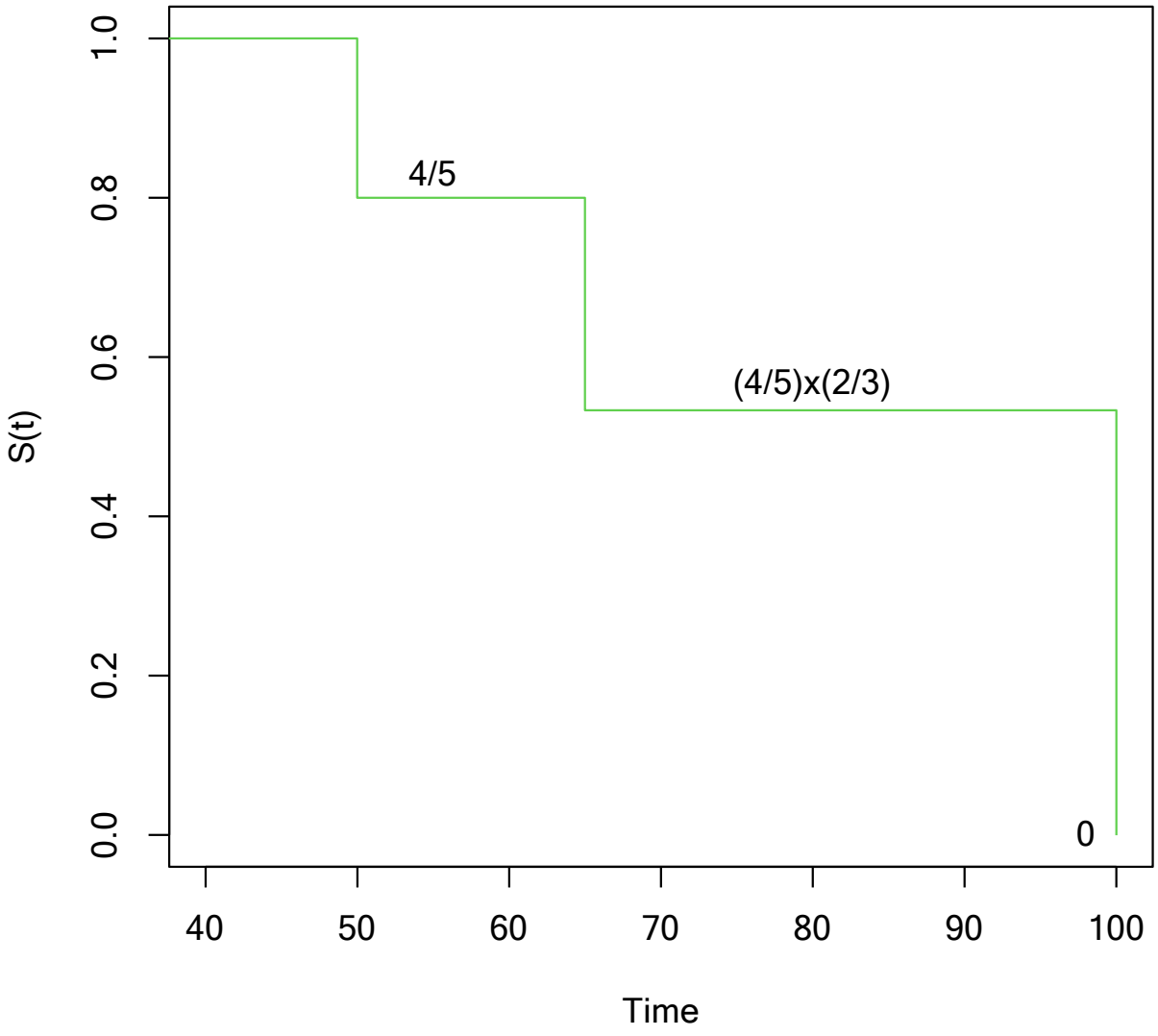
2番目の失敗生起事象



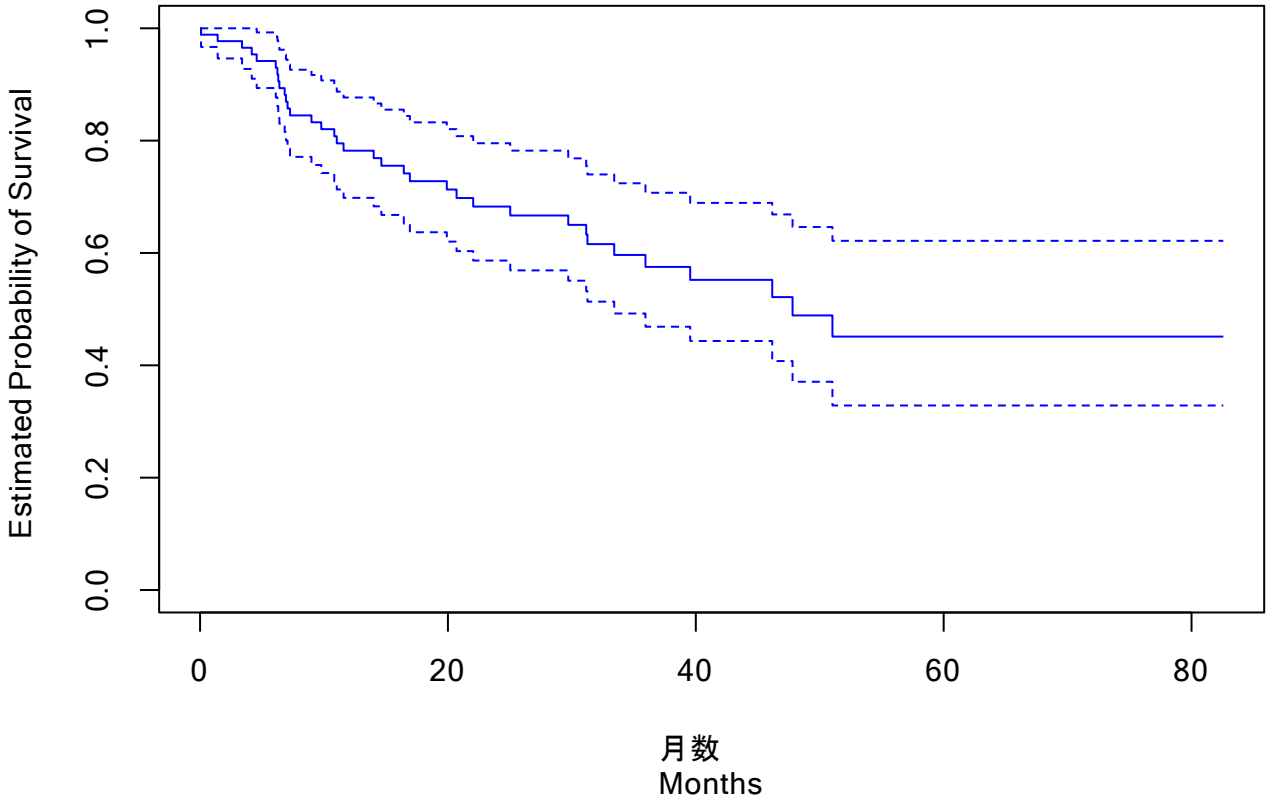
3番目の失敗生起事象



KM生存曲線

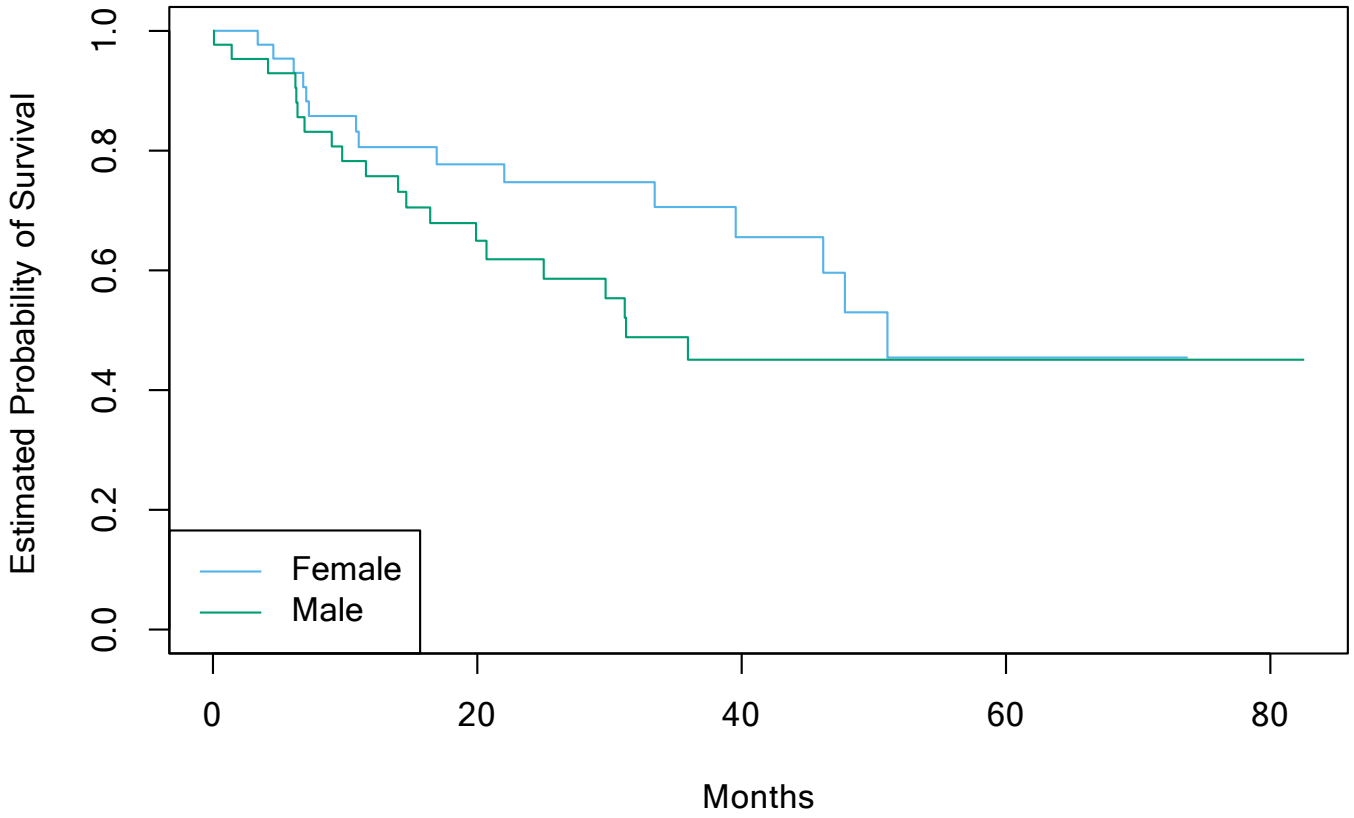


脳腫瘍データのKP(Kaplan-Meier)生存曲線



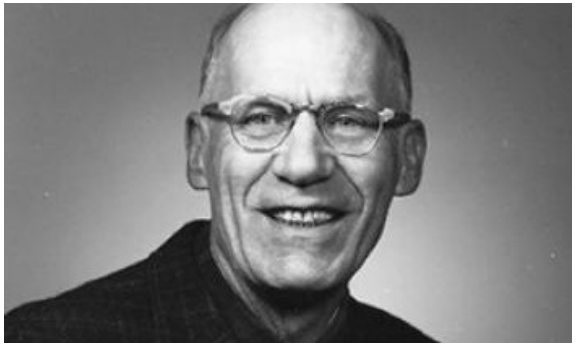
階段状の実線の点は横軸上の時刻を過ぎる生存確率を示す.20月以上の生存確率は71%,となり,以前に示したナイーブな推定値55% よりもかなり高い.

ログ・ランクテスト
-Log-Rank Test-

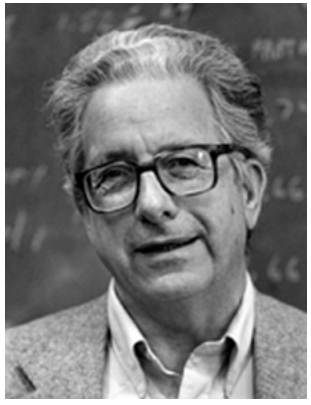


男性の生存時間と女性の生存時間を比較する．この二つのグループに関するKaplan-Meier生存曲線を示している．

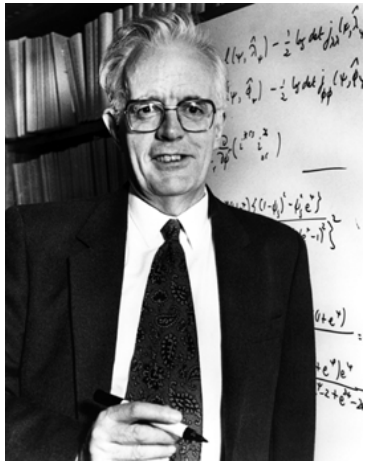
生存時間解析の主要な貢献者



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel **William Haenszel**

(***log rank test***)

Terry Therneau

(author of Survival package in R)

ログランク検定ー続き

- 50ヵ月ぐらいまでは女性の方が生存確率が高いと言えそうであるが, 50%程度で男女の曲線はカットされている．二つの生存曲線の同等性をどのように検定したら良いだろうか？

ログランク検定ー続き

- 50ヵ月ぐらいまでは女性の方が生存確率が高いと言えそうであるが, 50%程度で男女の曲線はカットされている. 二つの生存曲線の同等性をどのように検定したら良いだろうか?
- まず2標本t−検定がわかりやすい方法のように思える, しかし打ち切りが存在するのでそう単純ではない.
- この問題を解決するため, ログランク検定(log−rank test)を行おう.

ログランク検定ー続き

- ここで$d_1 < d_1 < \cdots < d_K$は打ち切られていない異なる死亡時刻, r_kは時刻d_k,におけるリスクを持つ患者数, q_kは時刻d_kにおいて死亡している患者数とする.

ログランク検定一続き

- ここで$d_1 < d_1 < \cdots < d_K$は打ち切られていない異なる死亡時刻, r_kは時刻d_k,におけるリスクを持つ患者数, q_kは時刻d_kにおいて死亡している患者数とする.
- さらにr_{1k}, r_{2k}を時刻d_kにおけるリスクを持つ患者グループ1, 2の患者数としよう.

ログランク検定一続き

- ここで$d_1 < d_1 < \cdots < d_K$は打ち切られていない異なる死亡時刻, r_kは時刻d_k,におけるリスクを持つ患者数, q_kは時刻d_kにおいて死亡している患者数とする.
- さらにr_{1k}, r_{2k}を時刻d_kにおけるリスクを持つ患者グループ1, 2の患者数としよう.
- 同様に, q_{1k}, q_{2k}を時刻d_kにおけるグループ1, 2での死亡した患者数としよう. ここで$r_{1k} + r_{2k} = r_k, q_{1k} + q_{2k} = q_k$である.

検定統計量

	Group 1	Group 2	合計
死亡	q_{1k}	q_{2k}	q_k
生存	$r_{1k} - q_{1k}$	$r_{2k} - q_{2k}$	$r_k - q_k$
合計	r_{1k}	r_{2k}	r_k

各死亡時刻d_kにおいて上のような2 × 2 のカウント表を作成.
ここで, 死亡時刻に重なりがなければ (すなわち同時刻に二名が亡くならない), q_{1k}と q_{2k}のどちらかが1, 他はゼロとなる.

ログランク検定:アイデア

- ある確率変数Xについて仮説H_0： $\mathrm{E}(X) = 0$を検定するには,
  一つの方法は次の形の検定統計量を構成:

$$W = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}$$

- ただし$\mathrm{E}(X)$, $\mathrm{Var}(X)$は仮説H_0の下でのXの期待値と分散である.

ログランク検定:アイデア

- ある確率変数Xについて仮説H_0： $\mathrm{E}(X) = 0$を検定するには,
  一つの方法は次の形の検定統計量を構成:

$$W = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}$$

- ただし$\mathrm{E}(X)$, $\mathrm{Var}(X)$は仮説H_0の下でのXの期待値と分散である.
- ログランク検定を行うには, まず次を求める. $X = \sum_{k=1}^{K} q_{1k}$, ただし$q_{1k}$は表の左上の数値で与えられる.

統計量

ログランク検定統計量は以下で与えられる

$$W = \frac{\sum_{k=1}^{K}\left(q_{1k} - \mathrm{E}(q_{1k})\right)}{\sqrt{\sum_{k=1}^{K}\mathrm{Var}(q_{1k})}} = \frac{\sum_{k=1}^{K}\left(q_{1k} - \frac{q_k}{r_k}r_{1k}\right)}{\sqrt{\sum_{k=1}^{K}\frac{q_k(r_{1k}/r_k)(1-r_{1k}/r_k)(r_k-q_k)}{r_k-1}}}.$$

標本数が大きければ, ログランク統計量Wは近似的に標準正規分布にしたがう.

このことを利用して帰無仮説：二つのグループ間で生存曲線に違いがない, に対してp−値が計算できる.

脳腫瘍データへの応用

- 脳腫瘍(BrainCancer)データにおける女性と男性の生存時間を比較すると, 統計量$W = 1.2$となり, 両側p−値は0.2となる.

脳腫瘍データへの応用

- 脳腫瘍(BrainCancer)データにおける女性と男性の生存時間を比較すると, 統計量$W = 1.2$となり, 両側p−値は0.2となる.
- したがって, 女性と男性で生存曲線に違いがないという帰無仮説は棄却できない.

脳腫瘍データへの応用

- 脳腫瘍(BrainCancer)データにおける女性と男性の生存時間を比較すると, 統計量$W = 1.2$となり, 両側p−値は0.2となる.
- したがって, 女性と男性で生存曲線に違いがないという帰無仮説は棄却できない.
- このログランク検定は次に述べるCoxの比例ハザードモデルに密接に関係する.

生存時間解析での回帰モデル

- 次に生存時間データに対して回帰モデルをフィットすることを考えよう.

生存時間解析での回帰モデル

- 次に生存時間データに対して回帰モデルをフィットすることを考えよう.
- 真の生存時間Tの予測を行いたい. ここで観測量$Y = \min(T, C)$は正値をとり, 右裾が厚い可能性があるので, $\log(Y)$をXに単回帰しようとするかもしれない. しかしながら,*打ち切りにより再び問題を生じる*.

生存時間解析での回帰モデル

- 次に生存時間データに対して回帰モデルをフィットすることを考えよう.
- 真の生存時間Tの予測を行いたい. ここで観測量$Y = \min(T, C)$は正値をとり, 右裾が厚い可能性があるので, $\log(Y)$をXに単回帰しようとするかもしれない. しかしながら,*打ち切りにより再び問題を生じる*.
- この問題に打ち勝つにはカプラン・マイヤー(Kaplan–Meiyer)生存曲線で利用したアイデアに類似する逐次的な方法を利用する.

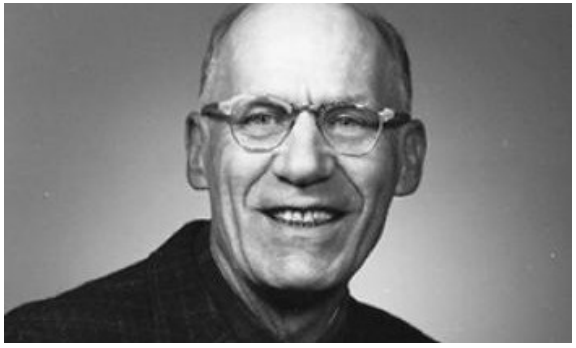
ハザード関数
Hazard Function

ハザード関数(*hazard function*), あるいはハザード率(*hazard rate*) ー 死力(force of mortality) ー は次式で定める

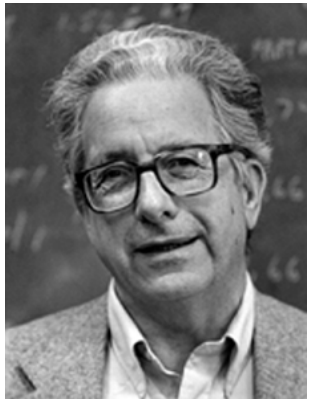
$$h(t) = \lim_{\Delta t \to 0} \Pr \frac{(t < T \leq t + \Delta t | T > t)}{\Delta t},$$

ただしTは(真の)生存時間である.
時刻tまでの生存を所与とするとその後には死力は一定である.
ハザード関数(hazard function)は次で述べる比例ハザードモデル(*Proportional Hazards Model*)の基礎となる.

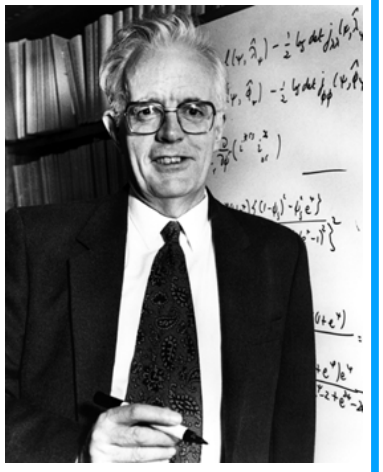
共変量covariatesの導入



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

William Haenszel

(*log rank test*)

Terry Therneau

(author of Survival package in R)

比例ハザード・モデル
-Proportional Hazards Model-

- 比例ハザードの仮定は次式で与えられる

$$h(t|x_i) = h_0(t)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right),$$

ここで$h_0(t) \geq 0$は特定化されないある関数, ベースライン・ハザード *(baseline hazard)*である. この関数は特徴量$x_{i1} = \cdots = x_{ip} = 0$となる個体のハザードを意味する.

比例ハザード・モデル
-Proportional Hazards Model-

- 比例ハザードの仮定は次式で与えられる

$$h(t|x_i) = h_0(t)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right),$$

ここで$h_0(t) \geq 0$は特定化されないある関数, ベースライン・ハザード *(baseline hazard)*である. この関数は特徴量$x_{i1} = \cdots = x_{ip} = 0$となる個体のハザードを意味する.
比例ハザード*(proportional hazards)*という名は特徴量ベクトルx_iの個体に対するハザードがある未知の関数$h_0(t)$に$\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$を乗ずることに起因する. 項$\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$はベクトル$x_i = (0, \ldots, 0)$に対する特徴量ベクトル$x_i = (x_{i1}, \ldots, x_{ip})$の相対リスク*(relative risk)*である.

比例ハザードモデル―続き

- ベースライン・ハザード関数$h_0(t)$を特定化しないとはどういう意味だろうか？

比例ハザードモデル―続き

- ベースライン・ハザード関数$h_0(t)$を特定化しないとはどういう意味だろうか？
- 基本的には, 関数形について何も仮定しない*(no assumptions about its functional form)*という意味である. 時刻tにおいてそれまで生存しているという条件の下で瞬間的な死亡確率はどのような形でもよい.

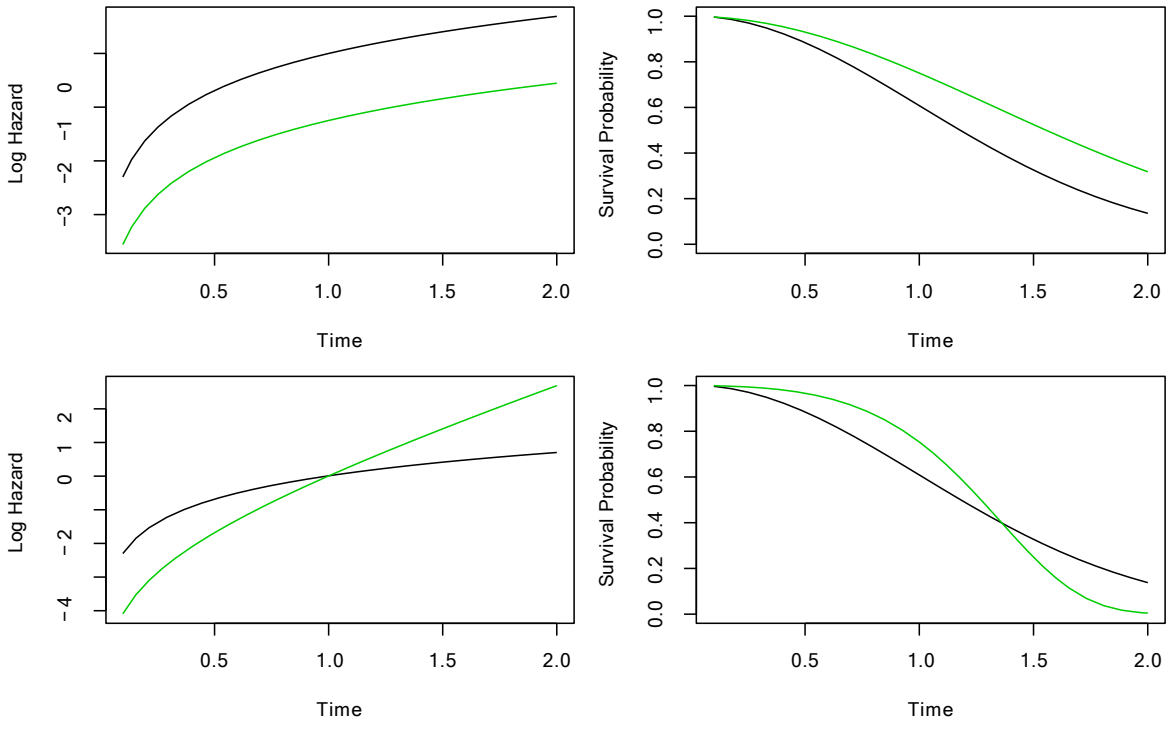
比例ハザードモデル―続き

- ベースライン・ハザード関数$h_0(t)$を特定化しないとはどういう意味だろうか？
- 基本的には，関数形について何も仮定しない*(no assumptions about its functional form)*という意味である．時刻tにおいてそれまで生存しているという条件の下で瞬間的な死亡確率はどのような形でもよい．
- このことはハザード関数は非常に柔軟，共変量(covariates)と生存時間のかなり広い関係のモデル化が可能となる．

比例ハザードモデル―続き

- ベースライン・ハザード関数$h_0(t)$を特定化しないとはどういう意味だろうか？
- 基本的には，関数形について何も仮定しない*(no assumptions about its functional form)*という意味である．時刻tにおいてそれまで生存しているという条件の下で瞬間的な死亡確率はどのような形でもよい．
- このことはハザード関数は非常に柔軟，共変量(covariates)と生存時間のかなり広い関係のモデル化が可能となる．
- ここでは，共変量x_{ij}の一単位変化により$\exp(\beta_j)$に対応する$h(t|x_i)$の変化に対応することのみを仮定している．

一例



$p = 1$, 2値共変量$x_i \in \{0,1\}$の例.

上図: 対数ハザードと生存関数 (グリーン$x_i = 0$, 黒$x_i = 1$). 比例ハザードの仮定からログハザード関数は定数のみ異なるので2つの生存時間関数は交わることがない.

下図: 比例ハザードの仮定を満たさない場合.

部分尤度−Partial Likelihood−

- ベースライン・ハザードの関数を未知としているので，尤度関数に$h(t|x_i)$を代入し，最尤推定法により母数$\beta = \left(\beta_1, \ldots, \beta_p\right)^T$を推定することはできない．

部分尤度-Partial Likelihood-

- ベースライン・ハザードの関数を未知としているので，尤度関数に$h(t|x_i)$を代入し，最尤推定法により母数$\beta = \left(\beta_1, \ldots, \beta_p\right)^T$を推定することはできない．
- Coxの比例ハザードモデルのマジックは$h_0(t)$の形を特定化することなく母数β*が推定できることである*．

部分尤度−Partial Likelihood−

- ベースライン・ハザードの関数を未知としているので, 尤度関数に$h(t|x_i)$を代入し, 最尤推定法により母数$\beta = \left(\beta_1, \ldots, \beta_p\right)^T$を推定することはできない.
- Coxの比例ハザードモデルのマジックは$h_0(t)$の形を特定化することなく母数β*が推定できることである*.
- その為にはKaplan−Meiyer法やログランク法で用いた時刻について連続的に利用するという論理を使う. 事象が起きた時刻y_iでのハザードは次で与えられる.

$$\sum_{i':y_{i'} \geq y_i} h_0(y_i) \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right).$$

部分尤度–続き

- したがって，第i番目の観測値が時刻y_iに失敗する確率(リスク集合にある他の観測値ではなく)は

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

部分尤度–続き

- したがって, 第i番目の観測値が時刻y_iに失敗する確率(リスク集合にある他の観測値ではなく)は

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

- ここでベースライン・ハザード関数$h_0(y_i)$は分母・分子でキャンセルする.

部分尤度– 続き

- 部分尤度は打ち切りのないすべての観測値についての確率の積で与えられ，

$$PL(\beta) = \prod_{i:\delta_i=1} \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

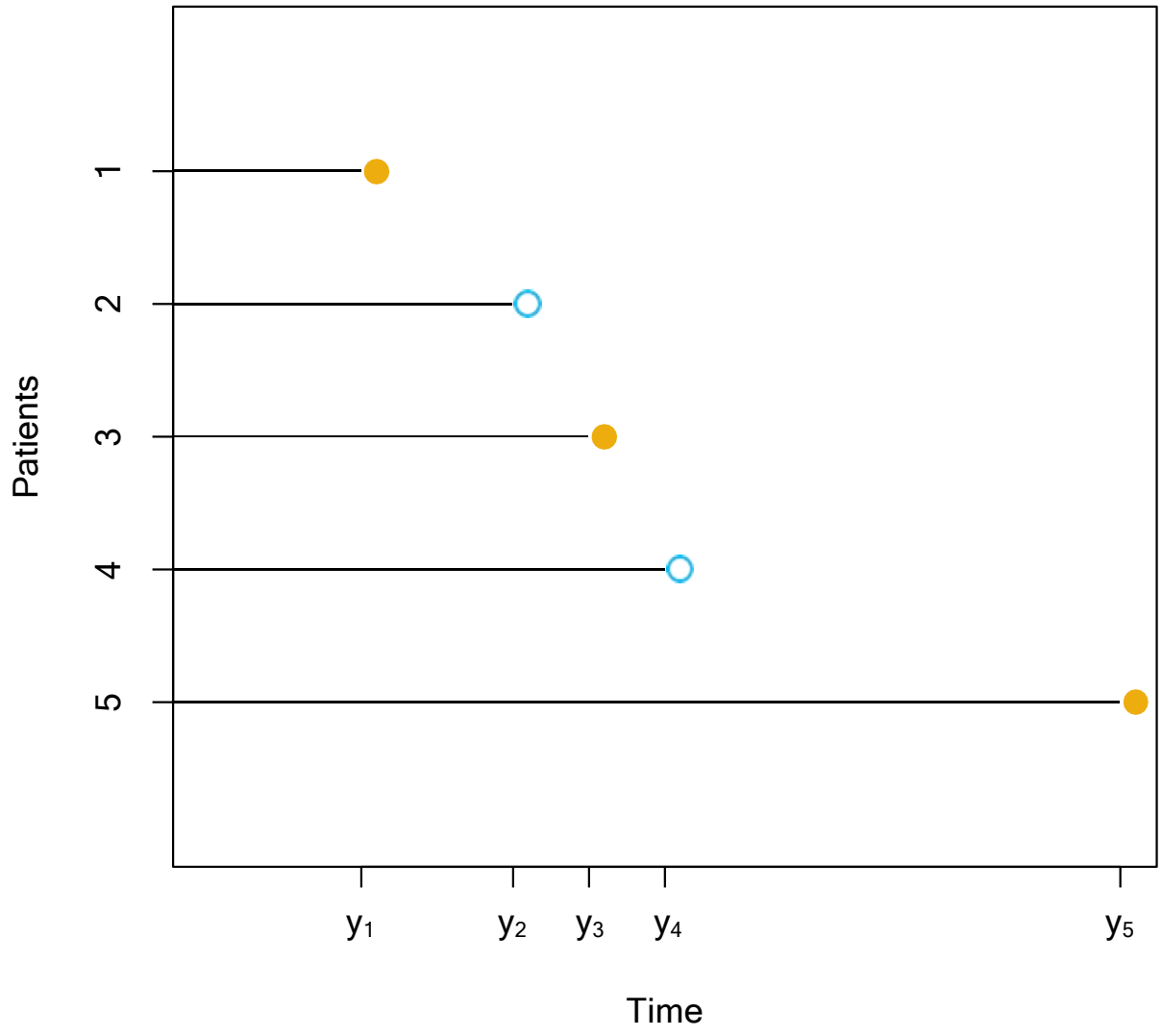
部分尤度– 続き

- 部分尤度は打ち切りのないすべての観測値についての確率の積で与えられ,

$$PL(\beta) = \prod_{i:\delta_i=1} \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

- 重要な点は部分尤度は真の$h_0(t)$の形に関わらず妥当であるので, モデルを柔軟で頑健なものにしてくれることだろう.

部分尤度: 例



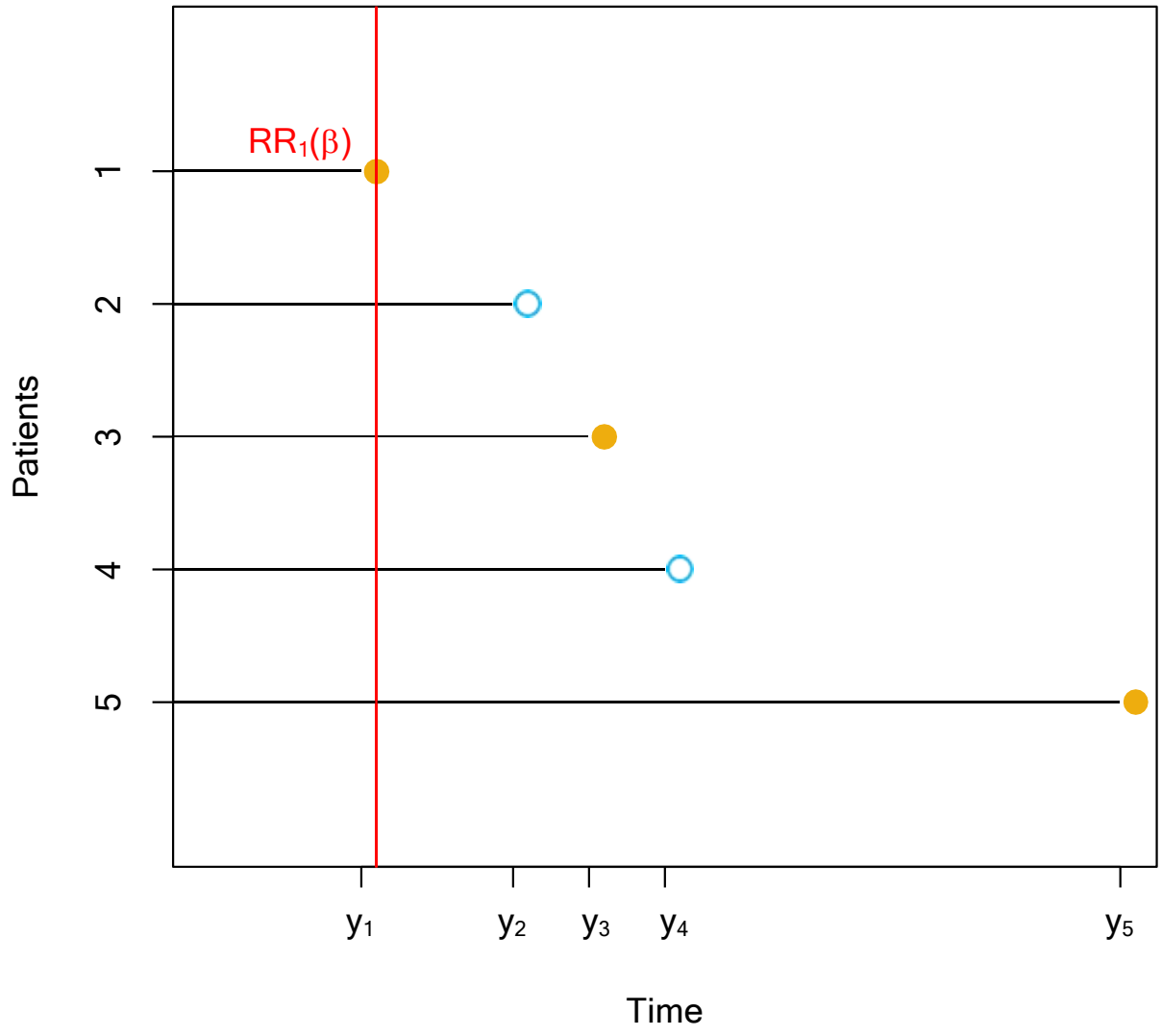
失敗生起時刻における相対リスク関数
Relative Risk Functions

$$RR_1(\beta) = \frac{\exp\left(\sum_{j=1}^{p} x_{1j}\beta_j\right)}{\sum_{i':y_{i'} \geq y_1} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

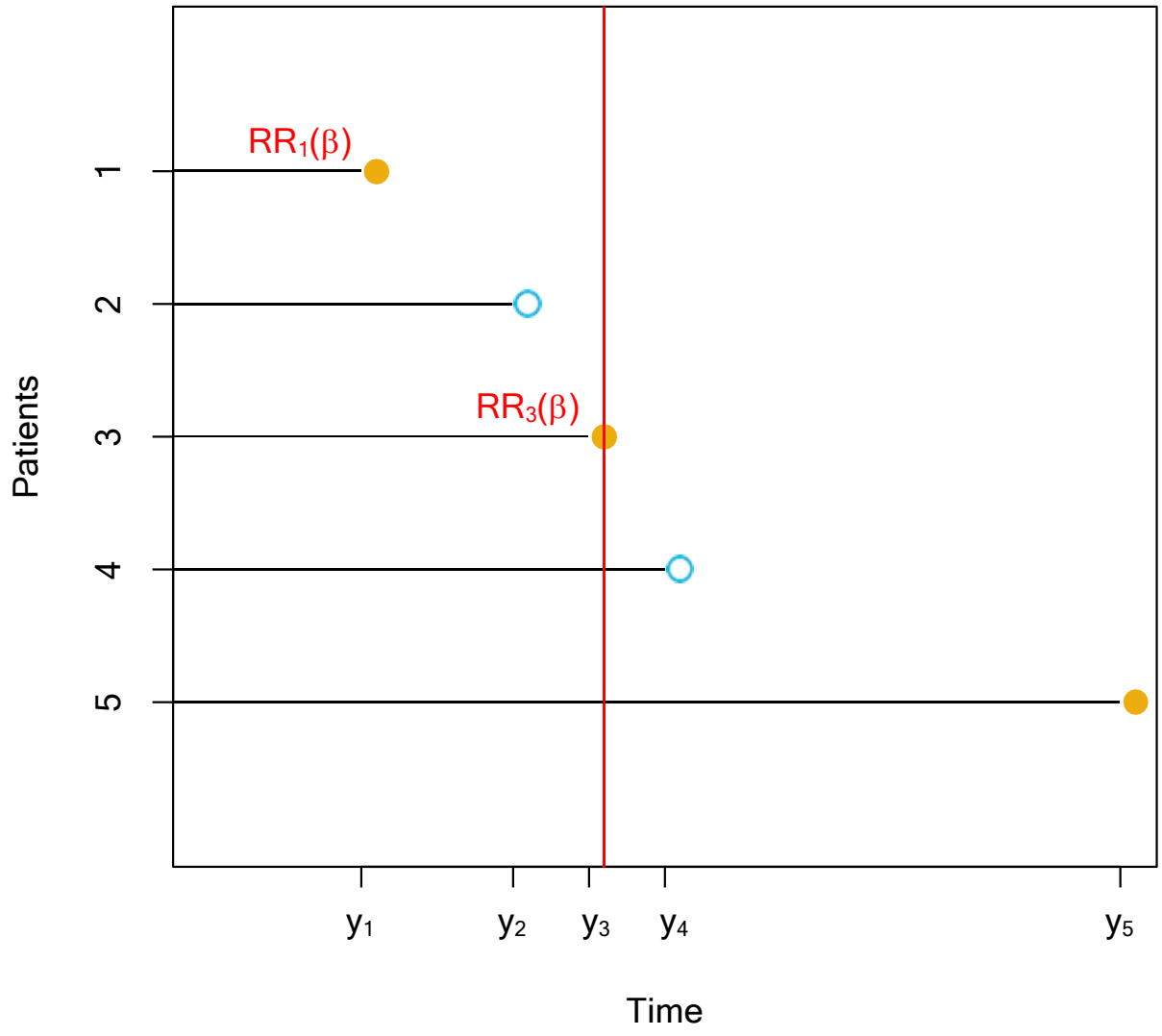
$$RR_3(\beta) = \frac{\exp\left(\sum_{j=1}^{p} x_{3j}\beta_j\right)}{\sum_{i':y_{i'} \geq y_3} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

$$RR_5(\beta) = \frac{\exp\left(\sum_{j=1}^{p} x_{5j}\beta_j\right)}{\sum_{i':y_{i'} \geq y_5} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}$$

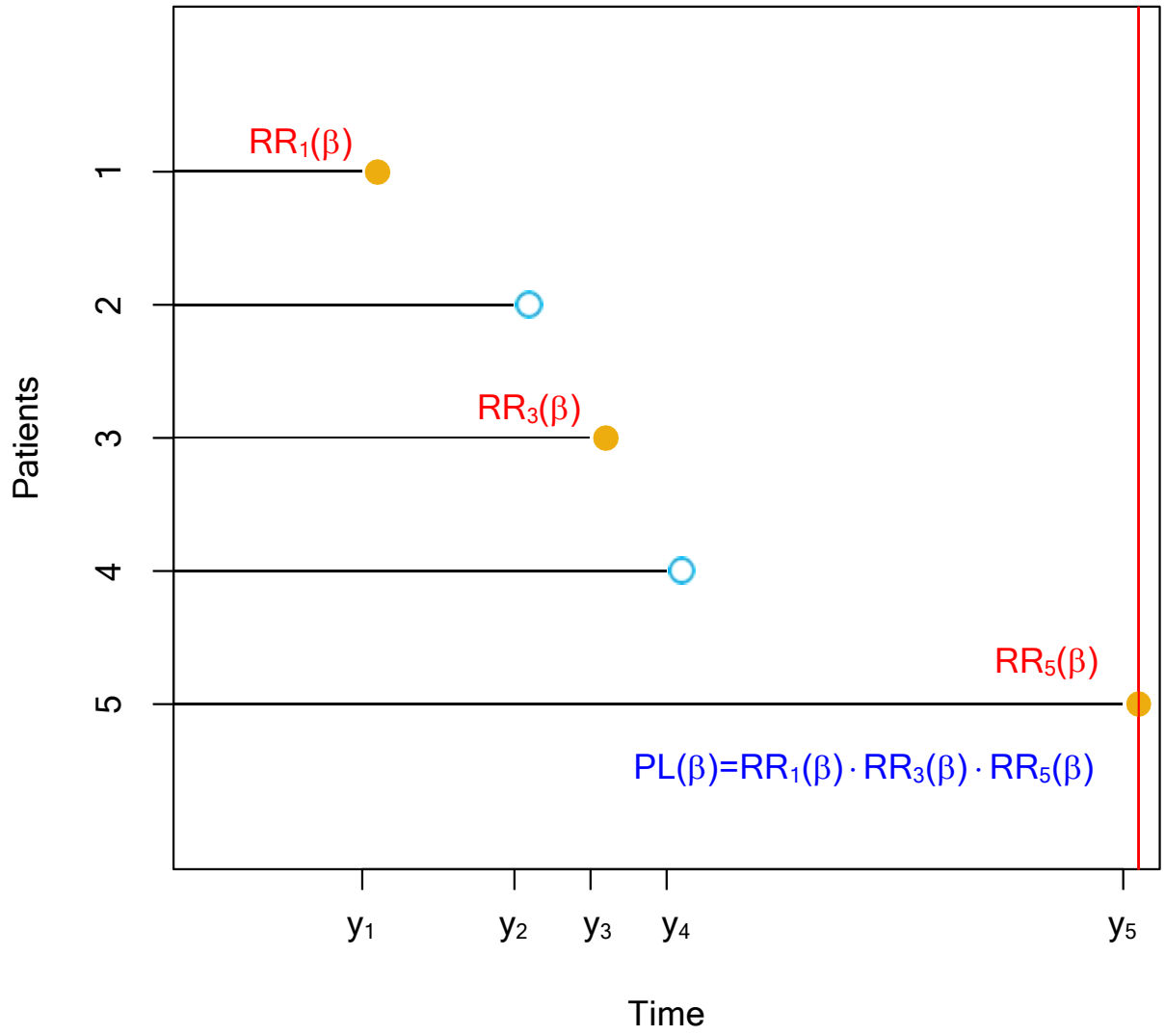
1番目の失敗生起事象



2番目の失敗生起事象



3番目の失敗生起事象


Patients
1
2
3
4
5
$RR_1(\beta)$
$RR_3(\beta)$
$RR_5(\beta)$
$PL(\beta)=RR_1(\beta)\cdot RR_3(\beta)\cdot RR_5(\beta)$
y_1
y_2
y_3
y_4
y_5
Time

部分尤度 – 計算

- 母数βを推定するにはβについて部分尤度を最大化すればよい. ロジスティック回帰モデルの場合には解は簡単な表現を持たないので繰り返しのアルゴリズムが必要となる.

部分尤度 – 計算

- 母数βを推定するにはβについて部分尤度を最大化すればよい. ロジスティック回帰モデルの場合には解は簡単な表現を持たないので繰り返しのアルゴリズムが必要となる.
- 母数βの推定に加えて, 最小二乗回帰やロジット回帰などと同様にモデル出力が得られる.

部分尤度 – 計算

- 母数βを推定するにはβについて部分尤度を最大化すればよい. ロジスティック回帰モデルの場合には解は簡単な表現を持たないので繰り返しのアルゴリズムが必要となる.
- 母数βの推定に加えて, 最小二乗回帰やロジット回帰などと同様にモデル出力が得られる.
- 例えば, 特定の帰無仮説($H_0: \beta_j = 0$など)に関するp−値を得たり, 係数の推定誤差や信頼区間などである.

ログランク検定との関係

- ここで1個の予測変数($p = 1$), $x_i \in \{0,1\}$とする. 2グループの観測値の生存時間に差があるか否かを検定するには二つの可能なアプローチが考えられる.

ログランク検定との関係

- ここで1個の予測変数($p = 1$), $x_i \in \{0,1\}$とする. 2グループの観測値の生存時間に差があるか否かを検定するには二つの可能なアプローチが考えられる.
    1. コックスの比例ハザードモデルをフィット, 帰無仮説 $H_0: \beta_j = 0$を検定する(p=1のときはβはスカラー.)

ログランク検定との関係

- ここで1個の予測変数($p = 1$), $x_i \in \{0,1\}$とする. 2グループの観測値の生存時間に差があるか否かを検定するには二つの可能なアプローチが考えられる.
    1. コックスの比例ハザードモデルをフィット, 帰無仮説 $H_0: \beta_j = 0$を検定する(p=1のときはβはスカラー.)
    2. 2群を比較するためログランク検定を行う.

ログランク検定との関係

- ここで1個の予測変数($p = 1$), $x_i \in \{0,1\}$とする. 2グループの観測値の生存時間に差があるか否かを検定するには二つの可能なアプローチが考えられる.
    1. コックスの比例ハザードモデルをフィット, 帰無仮説 $H_0: \beta_j = 0$を検定する(p=1のときはβはスカラー.)
    2. 2群を比較するためログランク検定を行う.

第一のアプローチのときには帰無仮説を検定するには幾つかの方法がある. 一つの方法はスコア法*score test*が知られている.

ログランク検定との関係

- ここで1個の予測変数($p = 1$), $x_i \in \{0,1\}$とする. 2グループの観測値の生存時間に差があるか否かを検定するには二つの可能なアプローチが考えられる.
    1. コックスの比例ハザードモデルをフィット, 帰無仮説$H_0: \beta_j = 0$を検定する(p=1のときはβはスカラー.)
    2. 2群を比較するためログランク検定を行う.

第一のアプローチのときには帰無仮説を検定するには幾つかの方法がある. 一つの方法はスコア法(*score test*)が知られている.
単一の2値共変量の場合にはコックスの比例ハザードモデルにおける帰無仮説$H_0: \beta_j = 0$に対するスコア統計量はログランク検定に正確に一致する.

比例ハザードモデルの追加事項

比例ハザードモデルの議論は幾つかの微妙な話題に関連している:

比例ハザードモデルの追加事項

比例ハザードモデルの議論は幾つかの微妙な話題に関連している:

- 比例ハザードモデルには切片がないが, 切片はベースラインハザードに吸収されている.

比例ハザードモデルの追加事項

比例ハザードモデルの議論は幾つかの微妙な話題に関連している:

- 比例ハザードモデルには切片がないが, 切片はベースラインハザードに吸収されている.
- これまで同時(タイ)の生存時間は存在しないことを仮定してきた. タイがあると部分尤度の正確な表現はより複雑化, 計算上で幾つかの近似を用いる必要がある.

比例ハザードモデルの追加事項

比例ハザードモデルの議論は幾つかの微妙な話題に関連している:

- 比例ハザードモデルには切片がないが, 切片はベースラインハザードに吸収されている.
- これまで同時(タイ)の生存時間は存在しないことを仮定してきた. タイがあると部分尤度の正確な表現はより複雑化, 計算上で幾つかの近似を用いる必要がある.
- 部分尤度*(partial likelihood)*は厳密な尤度ではない為の名前である. ただし近似としてはかなり良い.

比例ハザードモデルの追加事項

比例ハザードモデルの議論は幾つかの微妙な話題に関連している:

- 比例ハザードモデルには切片がないが, 切片はベースラインハザードに吸収されている.
- これまで同時(タイ)の生存時間は存在しないことを仮定してきた. タイがあると部分尤度の正確な表現はより複雑化, 計算上で幾つかの近似を用いる必要がある.
- 部分尤度*(partial likelihood)*は厳密な尤度ではない為の名前である. ただし近似としてはかなり良い.
- ここでは母数βの推定に焦点をあてている. ただしベースラインハザード $h_0(t)$の推定に関心があるかもしれない. 例えば生存曲線$S(t|x)$が推定できる. こうしたことは Rパッケージsurvival により実行できる.

例：脳腫瘍データ

変数	係数	se	z-統計量	p-値
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	<0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

訳注：医学専門用語については訳者など非専門家には正確な理解が困難である。例えばki (Karnofsky指標),Glioma (神経膠腫),Supertentorial (テント上), gtv(残存する浮腫)とは推測に過ぎない。

例：脳腫瘍データ

変数	係数	se	z-統計量	p-値
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	<0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

- この表は脳腫瘍データに比例ハザードモデルをフィットした結果を示している.

例：脳腫瘍データ

変数	係数	se	z-統計量	p-値
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	<0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

- この表は脳腫瘍データに比例ハザードモデルをフィットした結果を示している.
- 例えばKarnofsky indexの1単位増加は瞬間的に死亡する確率が$\exp(-0.05) = 0.95$の積を意味している.

例:脳腫瘍データ

変数	係数	se	z-統計量	p-値
sex[Male]	0.18	0.36	0.51	0.61
diagnosis[LG Glioma]	0.92	0.64	1.43	0.15
diagnosis[HG Glioma]	2.15	0.45	4.78	0.00
diagnosis[Other]	0.89	0.66	1.35	0.18
loc[Supratentorial]	0.44	0.70	0.63	0.53
ki	-0.05	0.02	-3.00	<0.01
gtv	0.03	0.02	1.54	0.12
stereo[SRT]	0.18	0.60	0.30	0.77

- この表は脳腫瘍データに比例ハザードモデルをフィットした結果を示している.
- 例えばKarnofsky indexの1単位増加は瞬間的に死亡する確率が$\exp(-0.05) = 0.95$の積を意味している.
- 言い換えると, Karnofsky indexが高ければ,任意の時刻における死亡するチャンスが低くなることを意味している. この効果は非常に有意でp値は0.0027となっている.

例: 出版データ

次に出版(Publication)データを扱うが, これは米国国立衛生研究所と血液研究所が出資した臨床試験の報告に関して学術誌での出版までに要した生存時間である.

例: 出版データ

次に出版(Publication)データを扱うが, これは米国国立衛生研究所と血液研究所が出資した臨床試験の報告に関して学術誌での出版までに要した生存時間である.

臨床試験244について出版までの月数を記録したデータである. 244の内, 研究期間内には156が出版されたが, 他の研究は打ち切られている.

例: 出版データ

次に出版(Publication)データを扱うが, これは米国国立衛生研究所と血液研究所が出資した臨床試験の報告に関して学術誌での出版までに要した生存時間である.

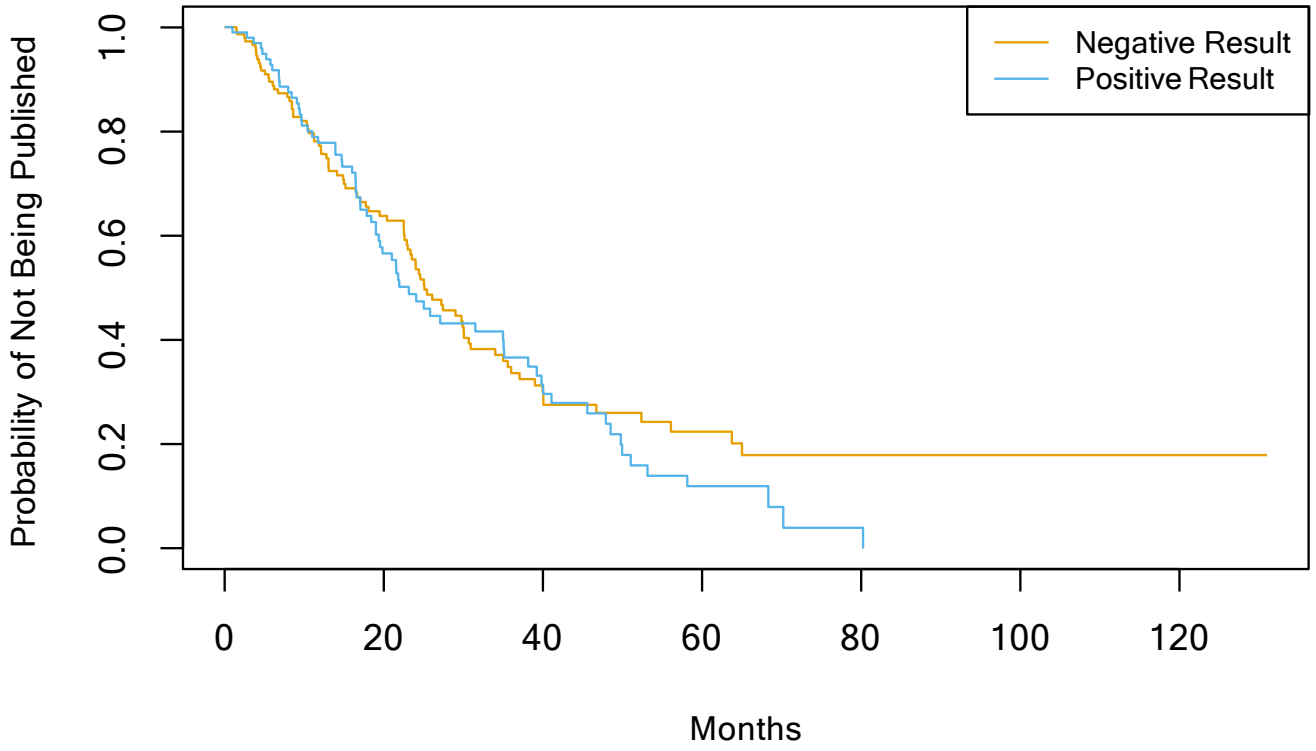
臨床試験244について出版までの月数を記録したデータである. 244の内, 研究期間内には156が出版されたが, 他の研究は打ち切られている.
共変量としては臨床試験のエンドポイントに焦点を当てた臨床試験か否か (clinend), 臨床試験が複数のセンターで行われたか否か (multi), 資金が国立衛生研究所の中か否か (mech), 臨床試験のサンプルサイズ (sampsize), 予算 (budget), インパクト(impact, 引用回数に関連), 臨床試験が正の(有意性)の結果をもたらしたか否か (posres)である.

例: 出版データ

次に出版(Publication)データを扱うが, これは米国国立衛生研究所と血液研究所が出資した臨床試験の報告に関して学術誌での出版までに要した生存時間である.

臨床試験244について出版までの月数を記録したデータである. 244の内, 研究期間内には156が出版されたが, 他の研究は打ち切られている.
共変量としては臨床試験のエンドポイントに焦点を当てた臨床試験か否か (clinend), 臨床試験が複数のセンターで行われたか否か (multi), 資金が国立衛生研究所の中か否か (mech), 臨床試験のサンプルサイズ (sampsize), 予算 (budget), インパクト(impact, 引用回数に関連), 臨床試験が正の(有意性)の結果をもたらしたか否か (posres)である.
最後の共変量は特に興味深い. と云うのは幾つかの研究では正の臨床試験の結果が高い出版の可能性があることを示しているからである.

出版データー続き



- 図は研究が正の結果をもたらしたか否かにより分割, 出版までのKaplan−Meiyer曲線を示している.
- 図は正の結果の研究の方が若干は低いことを示している.しかしログランク検定ではあまり有意でないp−値0.36であった.

出版データ:多変量解析

	係数	Std. error	z-統計量	p-値
posres[Yes]	0.55	0.18	3.02	0.00
multi[Yes]	0.15	0.31	0.47	0.64
clinend[Yes]	0.51	0.27	1.89	0.06
mech[K01]	1.05	1.06	1.00	0.32
幾つは省略				
sampsize	0.00	0.00	0.19	0.85
budget	0.00	0.00	1.67	0.09
impact	0.06	0.01	8.23	0.00

- すべての特徴量を利用してコックスの比例ハザードモデルを推定した結果を示しておく.
- 他の共変量を固定した時, 正の結果の研究の出版確率は$e^{0.55} = 1.74$倍, ネガティブな結果に比べて高い.
- 変数posresの非常に小さなp−値は結果が非常に有意であることを示している.

深掘り

問題を深掘りするため，次のスライドでは他の共変量を調整，正の結果と負の結果にもとづく生存時間曲線の推定値を示す

深掘り

問題を深掘りするため, 次のスライドでは他の共変量を調整, 正の結果と負の結果にもとづく生存時間曲線の推定値を示す

生存時間曲線を作成する為に, ベースライン・ハザード率$h_0(t)$を推定した: Rパッケージsurvivalを利用したが, その細部はこの授業では扱わない.

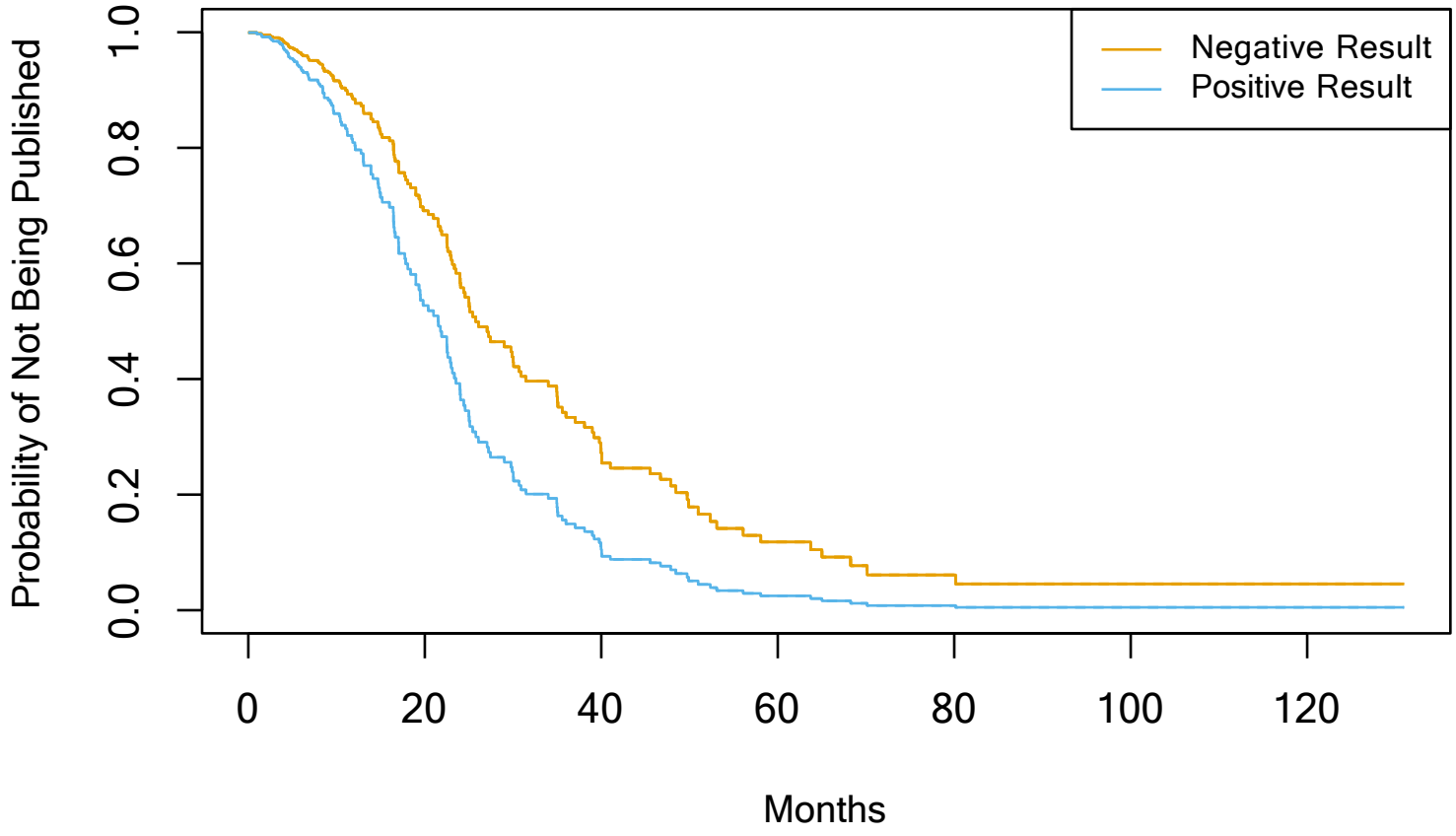
深掘り

問題を深掘りするため, 次のスライドでは他の共変量を調整, 正の結果と負の結果にもとづく生存時間曲線の推定値を示す

生存時間曲線を作成する為に, ベースライン・ハザード率$h_0(t)$を推定した: Rパッケージsurvivalを利用したが, その細部はこの授業では扱わない.

他の予測変数については代表値を選ぶ必要があるが, カテゴリカル予測変数mech (もっとも一般的なカテゴリー値R01)を除き, 各変数の平均値を利用した.

修正生存曲線
-Adjusted Survival Curves-



他の予測変数を調整すると, 正の結果と負の結果の研究における生存時間には明確な差が認められる. *[どうしてだろう ?]*

生存時間分析AUC: C−指標C−index

- テスト・データに対してコックス・モデルをフィットする為の有効な方法である.

生存時間分析AUC: C−指標C−index

- テスト・データに対してコックス・モデルをフィットする為の有効な方法である.
- 各観測値に対して推定したコックスモデルの係数を用いて推定したリスク・スコア$\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, (i = 1, \ldots, n)$を計算する.

生存時間分析AUC: C−指標C−index

- テスト・データに対してコックス・モデルをフィットする為の有効な方法である.
- 各観測値に対して推定したコックスモデルの係数を用いて推定したリスク・スコア$\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, (i = 1, \ldots, n)$を計算する.
- 次にHarrell'の一致指数(concordance index, *C−index*) を各観測値ペアについて計算$\hat{\eta}_{i'} > \hat{\eta}, y_i > y_{i'}$:

$$C = \frac{\sum_{i,i':y_i>y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i>y_{i'}} \delta_{i'}}.$$

生存時間分析AUC: C−指標C−index

- テスト・データに対してコックス・モデルをフィットする為の有効な方法である.
- 各観測値に対して推定したコックスモデルの係数を用いて推定したリスク・スコア$\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, (i = 1, \ldots, n)$を計算する.
- 次にHarrellの一致指数(concordance index, *C−index*) を各観測値ペアについて計算$\hat{\eta}_{i'} > \hat{\eta}, y_i > y_{i'}$:

$$C = \frac{\sum_{i,i':y_i>y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i>y_{i'}} \delta_{i'}}.$$

この値は生存時間が求まるすべてのペアの中で生存時間を正しく予測できる割合を意味する.

C–指標：例

出版データの訓練データにコックス比例ハザードモデルをフィットしてテストデータでC–指標を計算した.

結果は$C = 0.733$. 大まかにはこのテストデータにおける臨床試験の論文については統計モデルによりどちらが早く出版されるか73.3% の精度で予測されている.

その他の話題

- ここで教科書で述べている幾つかの追加的な話題に言及しておく.

その他の話題

- ここで教科書で述べている幾つかの追加的な話題に言及しておく.

他のタイプの打ち切り(censoring) : 左打ち切りと区間打ち切り(left and interval censoring).

その他の話題

- ここで教科書で述べている幾つかの追加的な話題に言及しておく.

他のタイプの打ち切り(censoring) : 左打ち切りと区間打ち切り(left and interval censoring).

時間軸の選択, 例えば, 暦上の時刻や年齢？

その他の話題

- ここで教科書で述べている幾つかの追加的な話題に言及しておく.

他のタイプの打ち切り(censoring) : 左打ち切りと区間打ち切り(left and interval censoring).

時間軸の選択, 例えば, 暦上の時刻や年齢？

時間に依存する共変量(*Time-dependent covariates*) ー 例えば異なる時刻における特徴量(例えば血圧)の計測値.

その他の話題

- ここで教科書で述べている幾つかの追加的な話題に言及しておく.

他のタイプの打ち切り(censoring) : 左打ち切りと区間打ち切り(left and interval censoring).

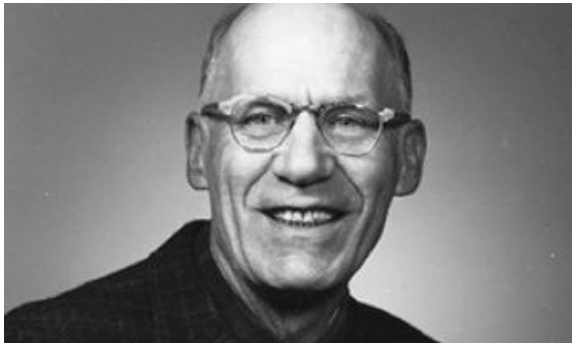
時間軸の選択, 例えば, 暦上の時刻や年齢 ?

時間に依存する共変量(*Time-dependent covariates*) ー 例えば異なる時刻における特徴量(例えば血圧)の計測値.

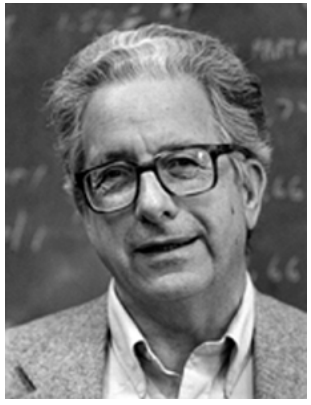
比例ハザードの仮定をチェックする方法.

他の機械学習法(*random forests*, *boosting* , *neural networks*など) を用いて生存時間データをモデル化する方法がある. 幾つかの方法は比例ハザードの仮定を避けている.

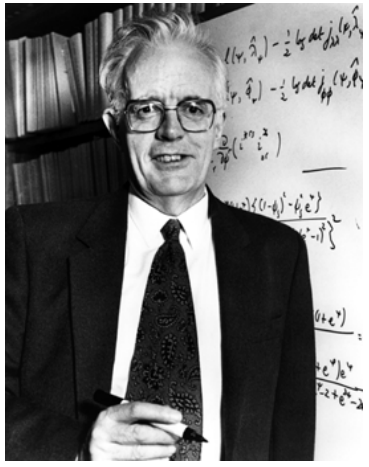
生存時間解析の主要な貢献者



Edward Kaplan



Paul Meier



David Cox

Nathan Mantel

William Haenszel

(***log rank test***)

Terry Therneau

(author of Survival package in R)

生存時間解析のソフトウエア

- この講義で取り上げた例はRパッケージsurvival,glmnetを用いて作られた.
- 二つのパッケージはともに時間に依存する共変量や一般の打ち切りを扱うことができる.

生存時間解析のソフトウエア

- この講義で取り上げた例はRパッケージsurvival,glmnetを用いて作られた.
- 二つのパッケージはともに時間に依存する共変量や一般の打ち切りを扱うことができる.
- 他の機械学習法についてのソフトウエアはR-repository, パイソンPythonソフトウエアscikit-survivalなどで利用できる.

第12章 : 教師なし学習
-Unsupervised Learning-

- 導入 : 教師あり学習と教師なし学習
- 主成分分析
- 犯罪データ
- K平均クラスタリング
- 階層的クラスタリング

第12章 : 教師なし学習
-Unsupervised Learning-

教師なし学習 対 教師あり学習:

- このコースの大部分は回帰や分類のような*教師あり学習*を扱っている.
- 教師あり学習の設定では、結果変数Yと特徴量$X_1, X_2, \dots, X_p$を観測している. $X_1, X_2, \dots, X_p$を用いてYを予測する事が目的となる.
- ここでは、*教師なし学習*、つまり$X_1, X_2, \dots, X_p$だけを観測した場合を扱う. 応答変数Yがないので、興味は予測ではなくなる.

教師なし学習の目的

- 目的は測定から得られる興味深い事を発見する事である. 例えば、データの情報を多く含んだまま可視化する方法はあるか？変数や観測の中に部分集団が見つけられるか？
- 2つの方法を議論する:
  - *主成分分析*, データの可視化や教師ありの方法を用いる前にデータの前処理に用いられる手法
  - *クラスタリング*, データの未知のサブグループを発見するための方法の広いクラスのこと

教師なし学習の困難さ

- 応答変数の予測のような単純な分析の目的がないため、教師なし学習は教師あり学習よりも主観的になる.
- しかし、教師なし学習の方法は多くの分野で重要度が増している:
  - 遺伝子表現の測定によって、乳がん患者からなるサブグループを見つける
  - 閲覧や購入の履歴から買い物客をグループ分けする
  - 映画の閲覧者による評価をもとに映画をグループ分けする

他の利点

- 実験装置やコンピュータから得られる*ラベルなしデータ*は、通常、人間の介入を必要とする*ラベルありデータ*よりも容易に得られる.
- 例えば、映画評価の総合的な感情、つまり好ましく思っているか否か、を自動的に評価する事は難しい.

主成分分析

- 主成分分析はデータセットの低次元での表現を与える.
- 教師あり学習の問題で用いるための変数を与えるためだけでなく、主成分分析はデータの可視化のツールとしての役割も果たす.

主成分分析: 詳細

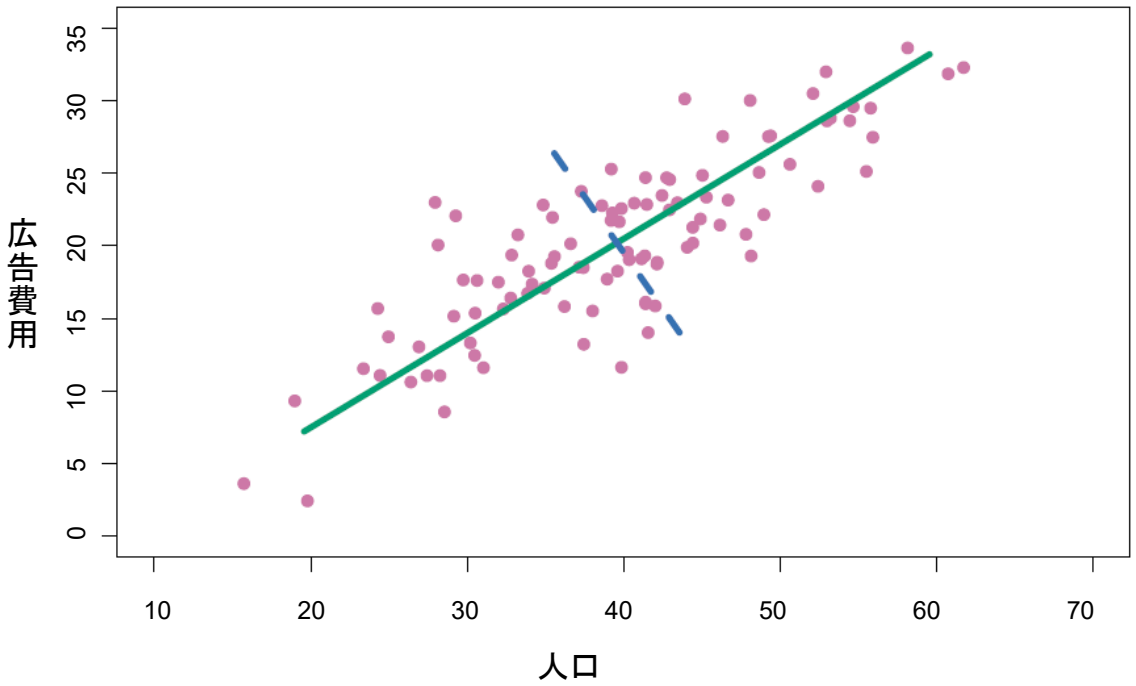
- 特徴量$X_1, X_2, \ldots, X_p$ の *第1主成分*は、正規化された線形結合で最大の分散を持つもの

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$

*正規化*は、$\sum_{j=1}^{p} \phi_{j1}^2 = 1$を意味する

- $\phi_{11} \ldots, \phi_{p1}$は第一主成分の負荷と呼ぶ. まとめて$\phi_1 = (\phi_{11} \ldots, \phi_{p1})^T$を負荷ベクトルと呼ぶ.
- 負荷の絶対値を大きくするといくらでも分散を大きくする事ができるため、負荷の2乗和が1に等しくなるように制約を置く.

主成分分析: 例



100の都市に関する、人口の大きさ(pop)と広告費用(ad)が点で示されている. 緑の実線は第一主成分の方向を表している. 青色の破線は第二主成分の方向を表している.

主成分の計算

- $n \times p$データセット$\boldsymbol{X}$があるとする. 分散にしか興味がないので、$\boldsymbol{X}$の各変数は平均0に中心化されている(つまり、Xの列平均は0)とする.
- $i = 1, \ldots, n$に対し、標本特徴量の線形結合を

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip} \qquad (1)$$

と表し、分散が最大となるものを$\sum_{j=1}^{p} \phi_{j1}^2 = 1$の制約の下で探す.

- x_{ij}の列平均が0なので、z_{i1}の平均も0. (ϕ_{j1}によらない.) よって、z_{i1}の標本分散は$\frac{1}{n}\sum_{i=1}^{n} z_{i1}^2$と表せる.

計算: 続き

- (1)を代入して、第一主成分の負荷ベクトルは、次の最適化問題の解.

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}\phi_{j1}^2 = 1\,.$$

- この問題は、$\boldsymbol{X}$の特異値分解によって解ける. 特異値分解は、線形代数の標準的手法である.
- $z_{11},\ldots,z_{n1}$を実現値とする、このZ_1を第一主成分と呼ぶ.

主成分の幾何

- $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$を成分とする負荷ベクトル$\phi_1$は特徴空間でデータが最も変動する方向を示す.
- n個のデータ点$x_1, \ldots, x_n$をこの方向に射影すると、その射影先の値は主成分スコア$z_{11}, \ldots, z_{n1}$になる.

さらなる主成分

- 第二主成分は、$X_1, \ldots, X_p$の線形結合で、Z_1と*無相関*なものの内、分散が最大となるものである.
- 第二主成分スコア$z_{12}, \ldots, z_{n2}$は

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots + \phi_{p2}x_{ip}$$

という形で、$\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$を成分とする第二主成分の負荷ベクトルを$\phi_2$で表す.

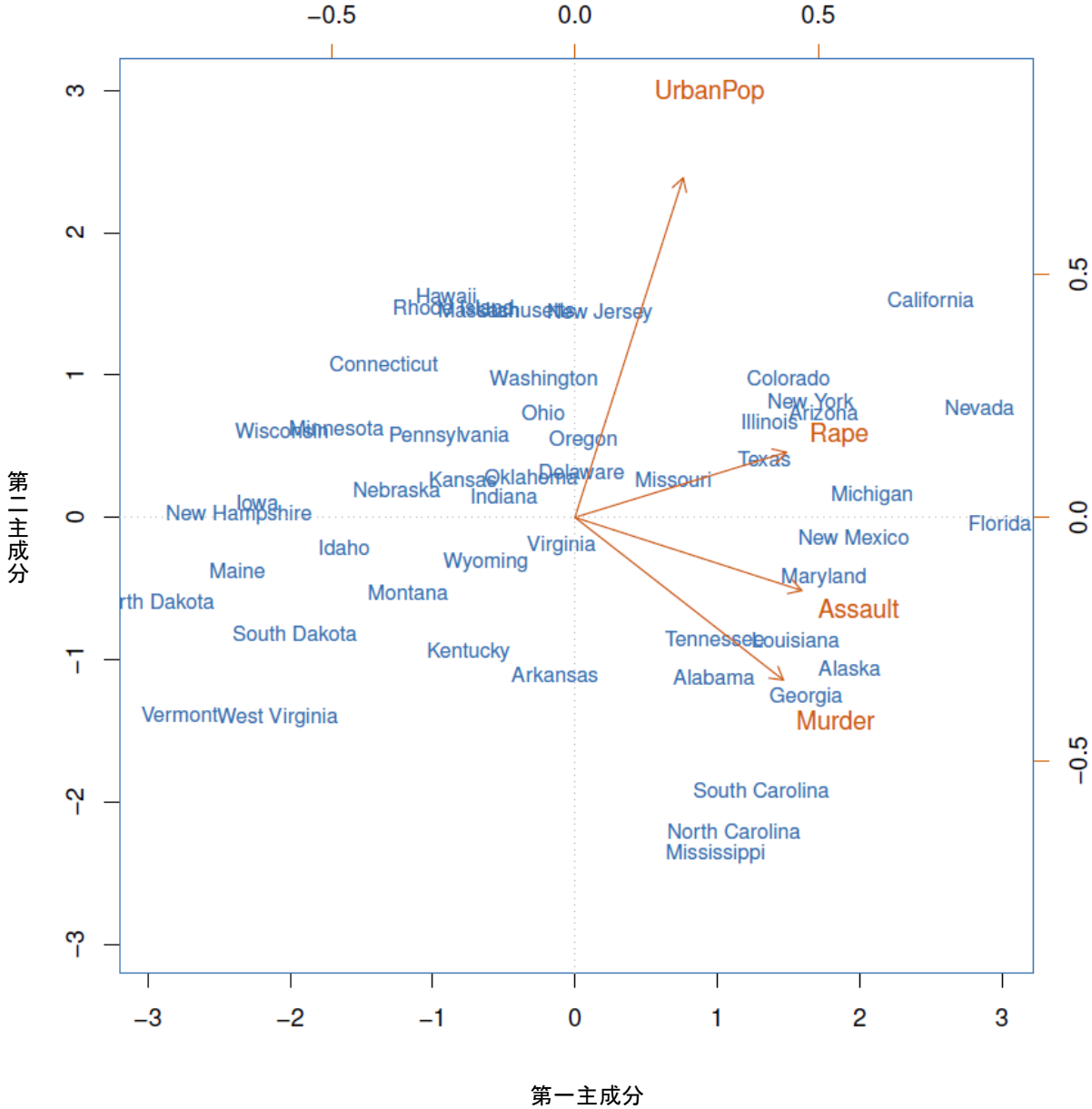
さらなる主成分: 続き

- Z_2がZ_1と無相関であるという制約は、ϕ_2がϕ_1と直交する事と等しい.
- 主成分の方向$\phi_1, \phi_2, \phi_3, \ldots$は行列$\boldsymbol{X}$の右特異ベクトルを順に並べたもので、各成分の分散は特異値の2乗の$\frac{1}{n}$倍である. 高々$\min(n-1, p)$個の主成分しかない.

例

- USAarrests(アメリカにおける犯罪)データ: アメリカの各50州に対し、Assault(暴行), Murder(殺人), Rape(強姦) の3つの犯罪に関する100, 000人当たりの逮捕者数を含むデータである. UrbanPop (各州で都市に住んでいる人の割合)も記録されている.
- 主成分スコアは長さ$n = 50$で、主成分の負荷ベクトルは長さ$p = 4$.
- 各変数の平均を0、標準偏差を1にする標準化した後で、主成分分析を行った.

アメリカにおける犯罪データ: 主成分分析のプロット



図の詳細

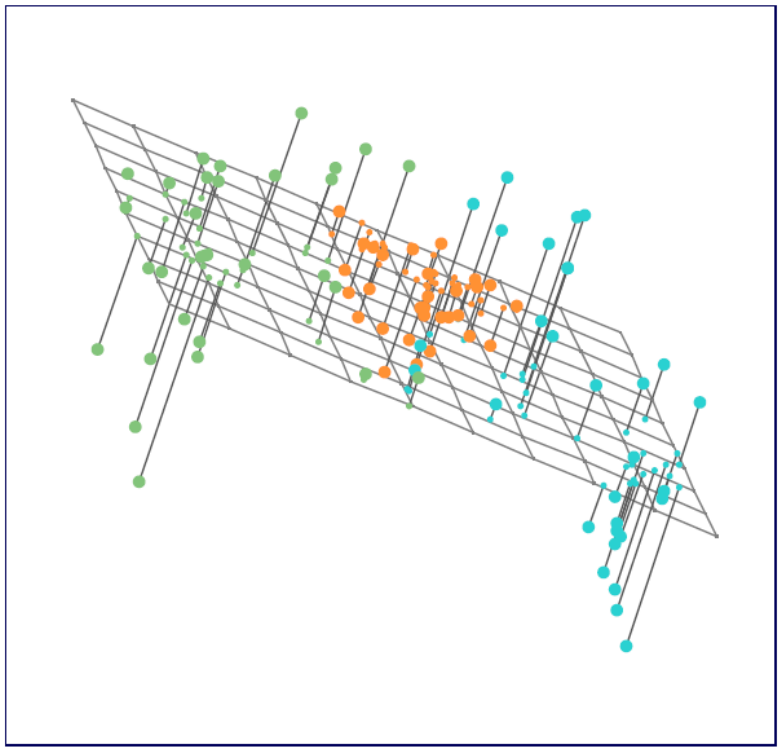
アメリカの犯罪データに関する初め2つの主成分.

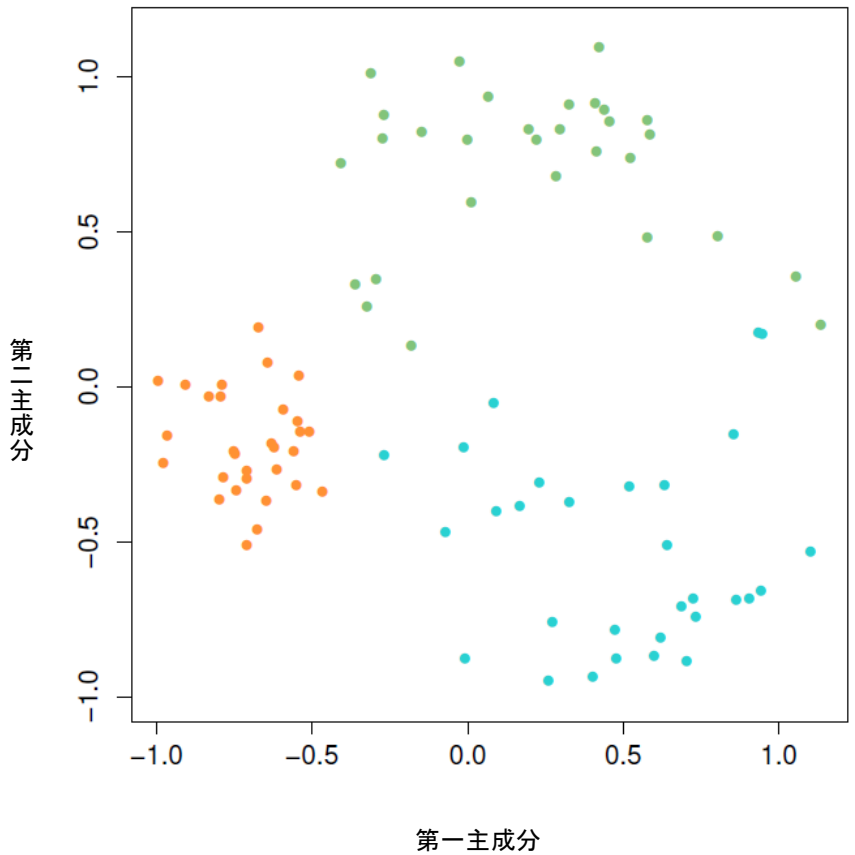
- 青色の州名は初め2つの主成分のスコアを表している.
- オレンジ色の矢印は初め2つの主成分の負荷ベクトルを表す(上と右に軸を持つ). 例えば、Rapeの第一主成分の負荷は0.54で、その第二主成分の負荷は 0.17 [Rapeは(0.54, 0.17)にある].
- この図は、主成分スコアと主成分の負荷を同時に図示するため、*バイプロット*と呼ばれる.

主成分分析 負荷

	PC1	PC2
Murder	0.5350995	-0.4181809
Assault	0.5031836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

主成分の他の解釈



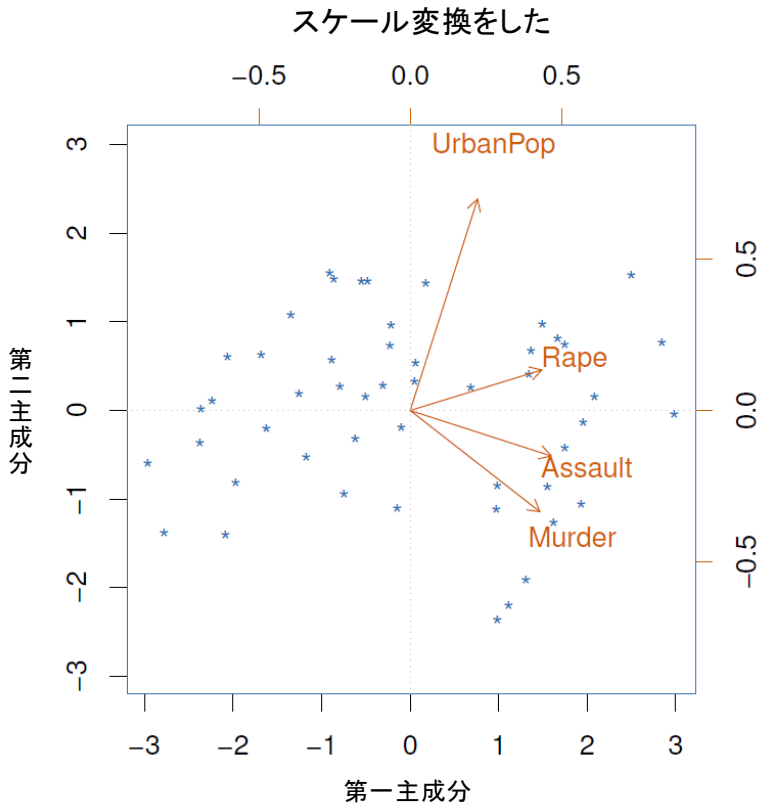


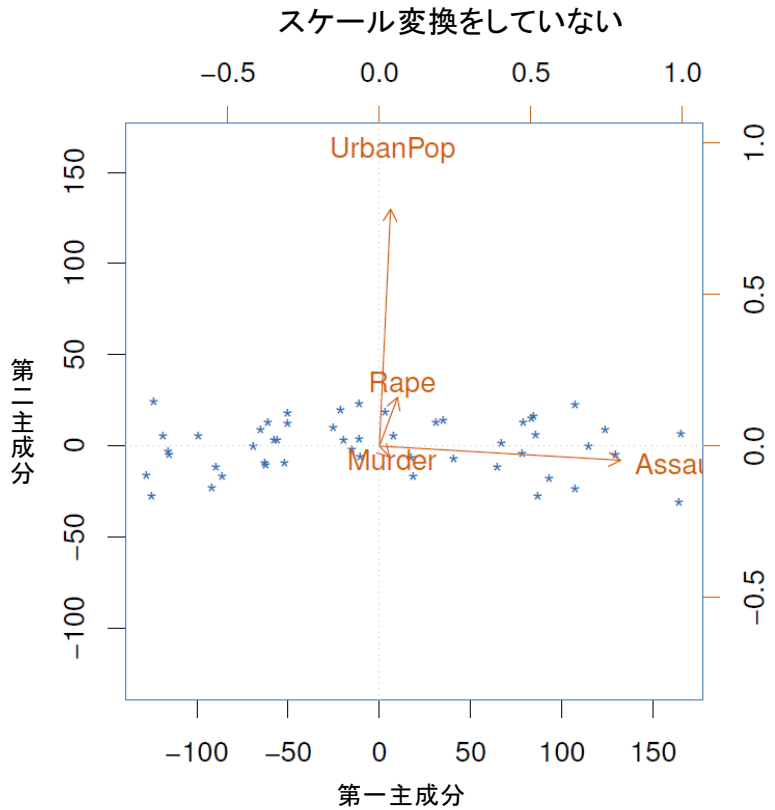
主成分分析は観測に最も近い超平面を見つける

- 第一主成分の負荷ベクトルはかなり特殊な性質を持っている: n個の観測に*最も近い*p次元空間上の線を定める(近さの指標としてはユークリッド距離による平均2乗距離を用いる.)
- n個の観測に最も近いという主成分の概念は、第一主成分だけでなく、各次数に拡張される.
- 例えば、データの初め2つの主成分がn個の観測に最も近い平面を張る. この距離もユークリッド距離による平均2乗距離による.

変数のスケールの問題

- 変数が異なる単位を持つとき、標準偏差が1となるようにスケール変換することが勧められる.
- 同じ単位を持つときは、スケール変換をしなくても良い場合とした方が良い場合がある.





説明済み分散の比率

- 各成分の大きさを理解するために、それぞれの説明済み分散の比率を知る事に興味がある.
- *総分散*は(中心化され平均が0である)データセットに対して

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

と定義され、第m主成分によって説明される分散は

$$\mathrm{Var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2 .$$

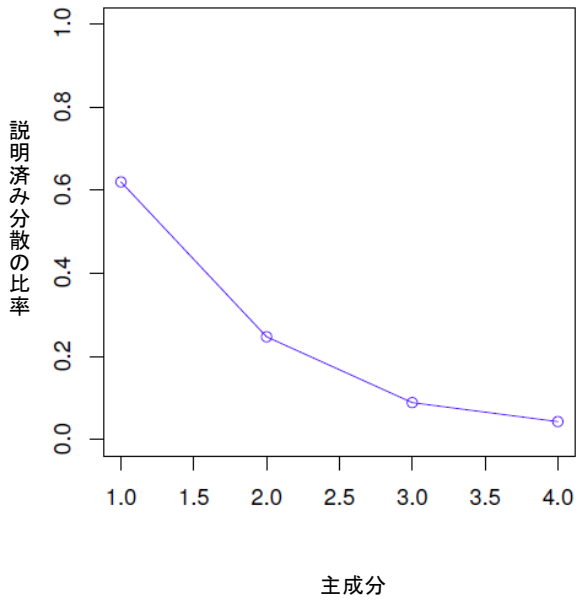
- $M = \min(n-1, p)$に対し、$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{m=1}^{M} \mathrm{Var}(Z_m)$であることが示せる.

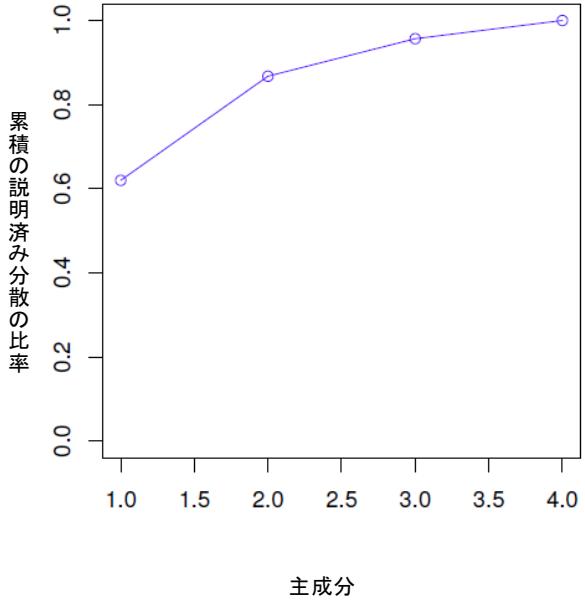
説明済み分散の比率: 続き

- よって、第m主成分の説明済み分散は0と1の間の正の値で

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}^2}.$$

- 説明済み分散を足し上げると1になる. 累積の説明済み分散を図示することがよくある.





何個の主成分を用いるべきか？

データの要約として主成分を用いるとき、何個の主成分を用いれば十分か？

- この問いへの簡単な答えはなく、交差検証を用いる事もできない.
  - *何故か？*

何個の主成分を用いるべきか？

データの要約として主成分を用いるとき、何個の主成分を用いれば十分か？

- この問いへの簡単な答えはなく、交差検証を用いる事もできない.
  - *何故か？*
  - どの様な時なら、交差検証を用いて主成分の数を選ぶ事が出来るか？

何個の主成分を用いるべきか？

データの要約として主成分を用いるとき、何個の主成分を用いれば十分か？

- この問いへの簡単な答えはなく、交差検証を用いる事もできない.
  - *何故か？*
  - どの様な時なら、交差検証を用いて主成分の数を選ぶ事が出来るか？
- 前のスライドの「スクリープロット」を指標として用いる事が出来る.「肘」を見つける.

クラスタリング

- *クラスタリング*は、あるデータの中の*サブグループ*や*クラスター*を見つけるためのかなり多くのテクニックのこと.
- 各グループの観測同士がかなり似るように、データをかぶりなくグループに分けたい.
- 問題を具体的にするためには、2つやそれ以上の観測が*似ている*又は*異なる*ということが何を意味するのかを決めなければならない.
- 実際、これは研究されているデータに関する知識に基づいた、分野特有の思考が必要となる.

主成分分析 対 クラスタリング

- 主成分分析は、分散を最も良く説明する、観測の低次元空間での表現を探す.
- クラスタリングは観測の中で同質なサブグループを探す.

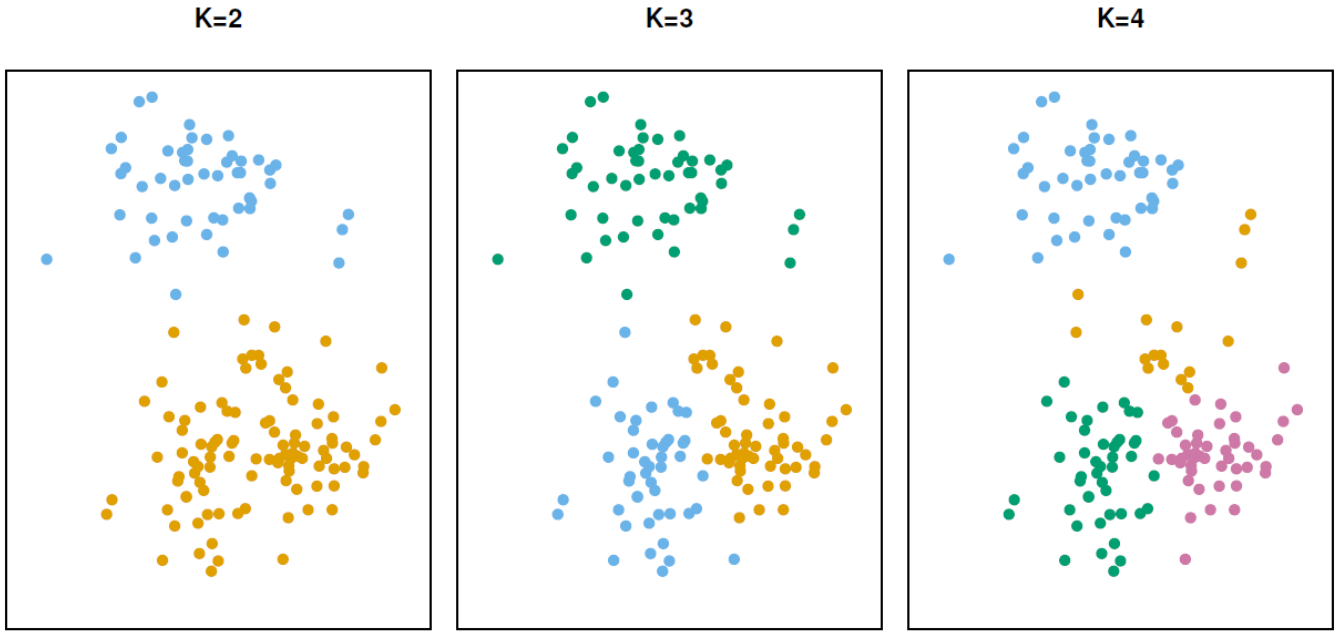
市場の細分化に対するクラスタリング

- 多くの人に対し、多くのデータ(例えば、世帯収入の中央値、職業、最も近い都市部からの距離、等).
- 目的は、ある製品を購入しそうな人々のサブグループや、ある広告に反応しやすい人々のサブグループを特定することによって、*市場の細分化*を行うことである.
- 市場の細分化を行う事は、データで人々をクラスタリングすることに相当する.

2つのクラスタリング法

- *K平均クラスタリング*では、観測を前もって定めた数のクラスタに分ける.
- *階層的クラスタリング*では、前もってはクラスタの数を知らない. *樹形図*と呼ばれる、木のような視覚化を得る. 1からnまでのあり得るクラスタの数に対してクラスタリングを得る事が出来る.

K平均クラスタリング



2次元空間上の、150個の観測のシミュレーションデータ. クラスタの数Kを変えて、K平均クラスタリングを適用した結果を示している. 各観測の色はK平均クラスタリングによって振り分けられたクラスタを示している. クラスタの順番はないので、クラスタの色はてきとう. クラスタのラベルはクラスタリングでは用いられておらず、クラスタリングの結果として得られるものである.

K平均クラスタリングの詳細

$C_1, \ldots, C_K$によって各クラスタの観測のインデックスの集合を表す:

1. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$. つまり、各観測は$K$個のクラスタのどれか1つに属する.
2. $k \neq k'$ に対し $C_k \cap C_{k'} = \emptyset$. つまり、クラスタに重なりがなく、2つ以上のクラスタに属する観測はない.

例えば、i番目の観測がk番目のクラスタに属するとき、$i \in C_k$と表す.

K平均クラスタリングの詳細: 続き

- K平均クラスタリングの背後にあるアイデアは*良い*クラスタリングを、*クラスタ内の分散*が出来るだけ小さくなるものだと考えること.
- クラスタC_kのクラスタ内分散は、クラスタ内の異なる観測の違いによる量$\mathrm{WCV}(C_k)$によって測る.
- よって、次の最小化問題を解きたい

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} WCV(C_k) \right\}. \tag{2}$$

- つまり、この定式化は、K個のクラスタ内分散を足したものが出来るだけ小さくなるように、観測を分けている.

クラスタ内分散をどのように定めるか

- ユークリッド距離を用いる事がよくある.

$$\mathrm{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2, \qquad (3)$$

ただし、$|C_k|$はk番目のクラスタの観測数を表す.

- (2)と(3)を合わせて、K平均クラスタリングを与える最適化問題は次のよう.

$$\underset{C_1,\ldots,C_K}{\mathrm{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2 \right\}. \qquad (4)$$

K平均クラスタリング アルゴリズム

1. 各観測に対して、1からKまでの数をランダムに割り当てる. これらがクラスタの割り付けの初期値となる.
2. 変化が止まるまでクラスタの割り付けを繰り返す:
   1. K個のクラスタそれぞれに対し、*重心*を計算する. k番目のクラスタの重心はk番目のクラスタの観測の平均によるp次元の特徴量ベクトル.
   2. 各観測を最も近い重心のクラスタに割り当てる (ただし*最も近い*とはユークリッド距離による).

アルゴリズムの性質

- (4)の目的関数の値が減少することが保証される. *何故か？*

アルゴリズムの性質

- (4)の目的関数の値が減少することが保証される. *何故か？*

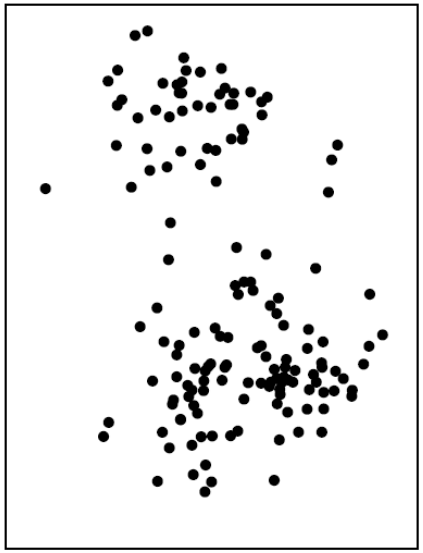
$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2,$$

ただし、$\bar{x}_{kj}=\frac{1}{|C_k|}\sum_{i\in C_k}x_{ij}$は$k$番目のクラスタの特徴$j$の平均を表す.

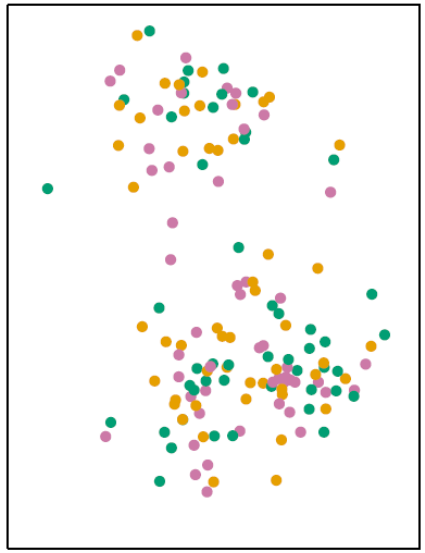
- しかし、大域局所解を与えるとは限らない. *何故か？*

例

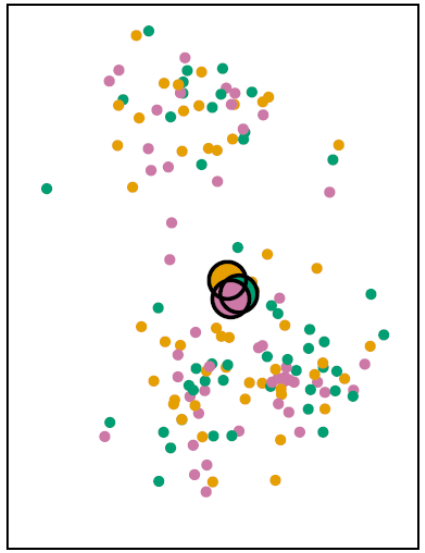
データ



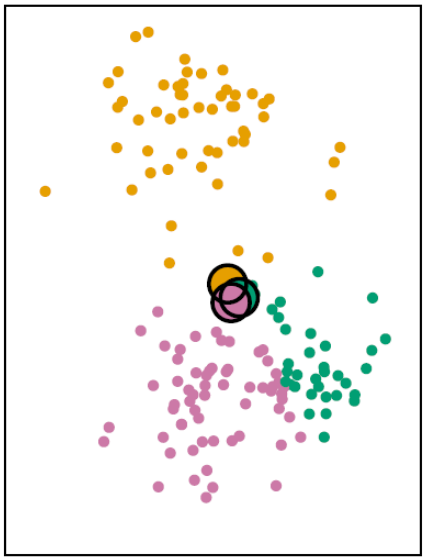
ステップ1



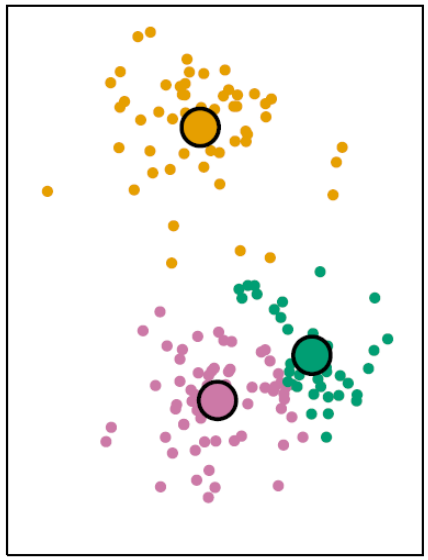
繰り返し1回目, ステップ2a



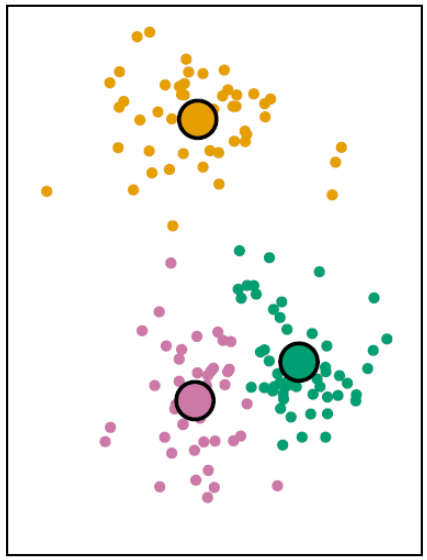
繰り返し1回目, ステップ2a



繰り返し2回目, ステップ2a



最終結果

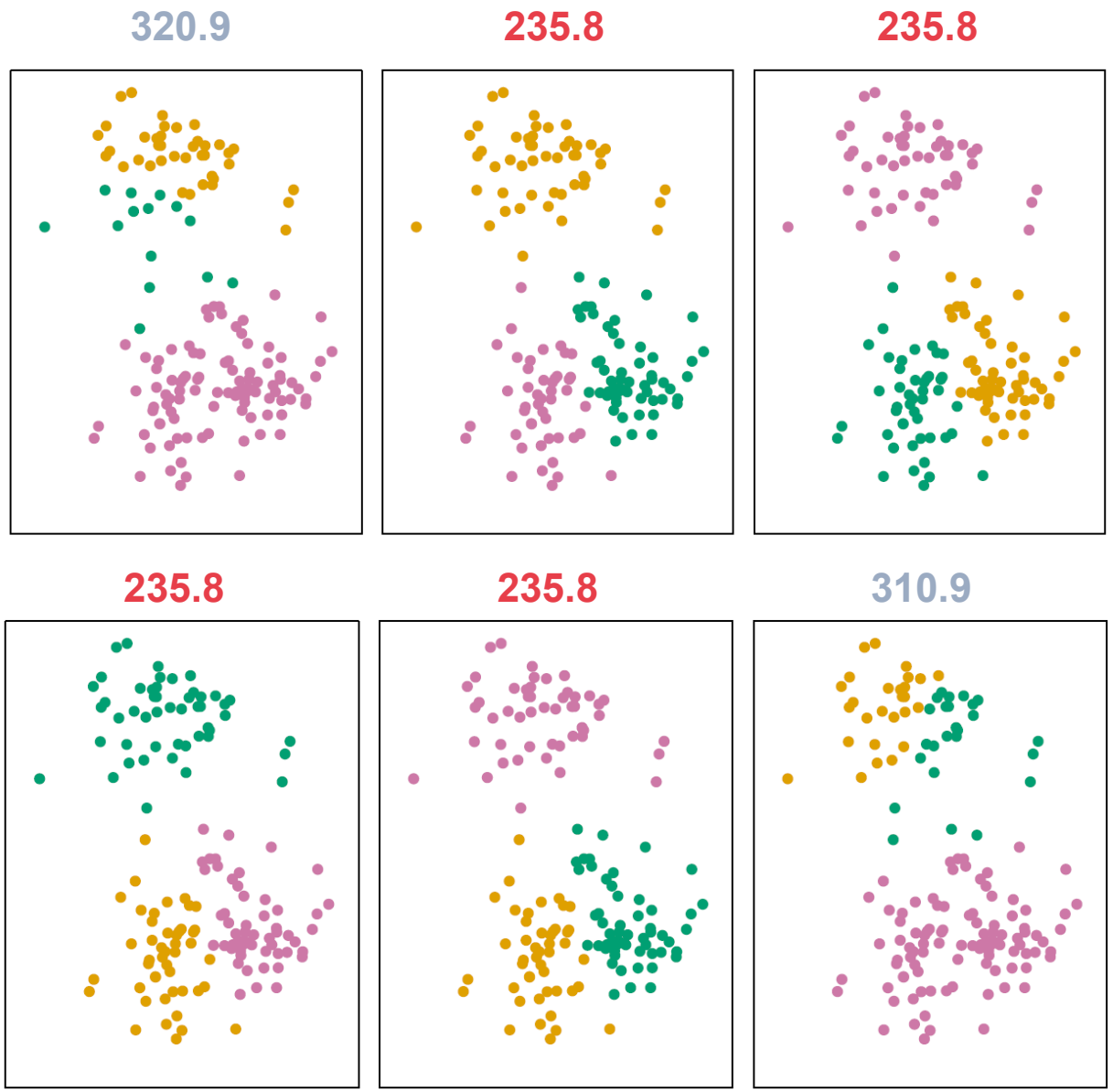


前の図の詳細

$K = 3$でのK平均アルゴリズムの進捗.

- *左上:* 観測が与えられている.
- *真ん中上:* ステップ1で、観測がランダムに割り付けられる.
- *右上:* ステップ2.1で、重心が計算される. これらは色付きの大きな円で表されている. データをランダムに割り付けているので、重心はほとんど重なっている.
- *左下:* ステップ2.2で、各観測が最も近い重心に割り付けられる.
- *真ん中下:* ステップ2.1が再度行われ、新たな重心を得る.
- *右下:* 10回の繰り返しの後に得られた結果.

例: 異なる初期値



前の図の詳細

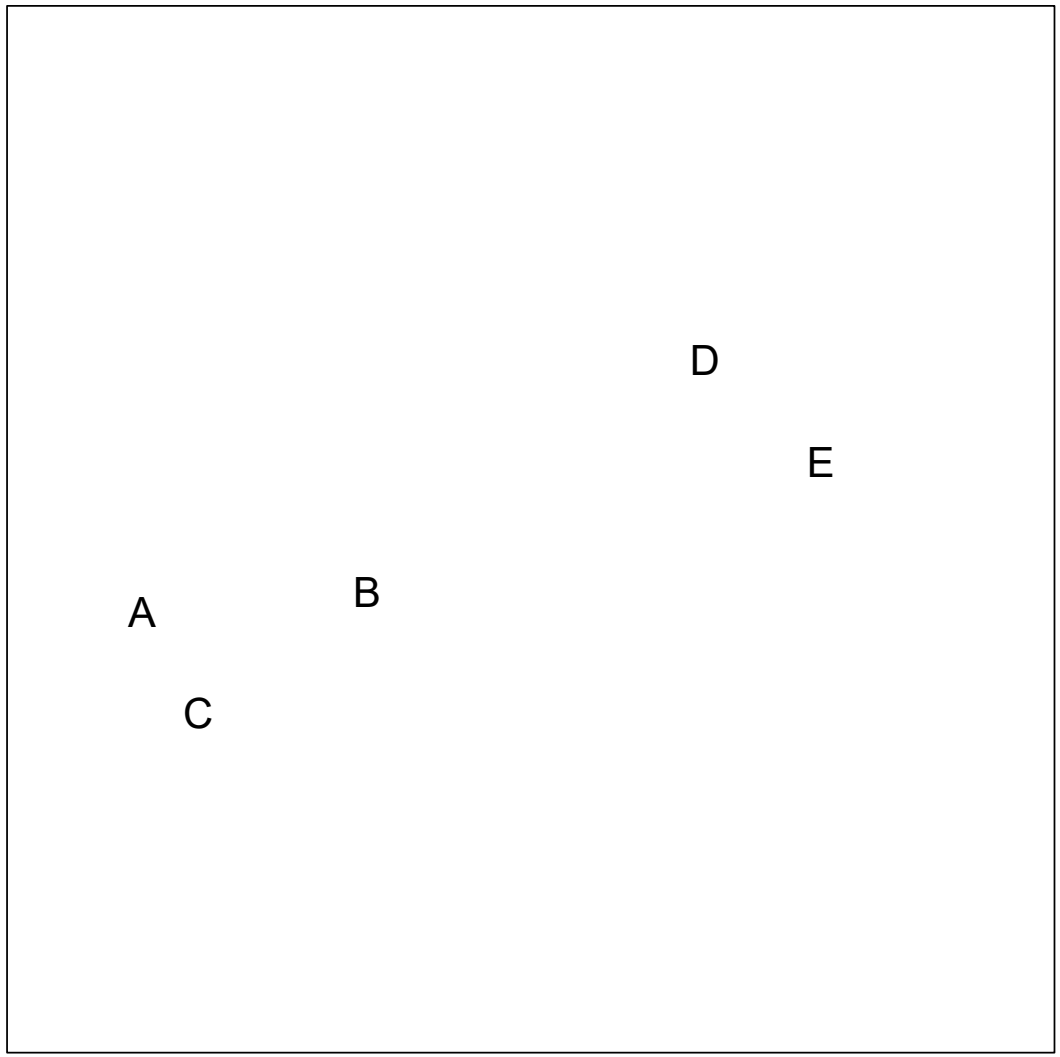
K平均クラスタリングを$K = 3$で先ほどのデータに対して6回行った. それぞれK平均アルゴリズムのステップ1での観測のランダムな割り付けが異なる.
各プロットの上の値は、目的関数(4)の値.
3つの異なる局所解が得られた. 目的関数をより小さくするものは、クラスタの分離を上手く行っている.
赤色になっている所では、すべて同じ最も良い解を得た. 目的関数の値は235.8である.

階層的クラスタリング

- K平均クラスタリングでは、前もってクラスタ数Kを定める必要があった(あとでKを選ぶ方法を議論する)
- *階層的クラスタリング*はKの選択をあらかじめ行う必要はない、別の方法である.
- このセクションでは、*ボトムアップ*又は*凝集*クラスタリングを説明する. これは階層的クラスタリングのもっともよく用いられるもので、葉から始まってクラスタを結合して幹にして、樹形図が作られる.

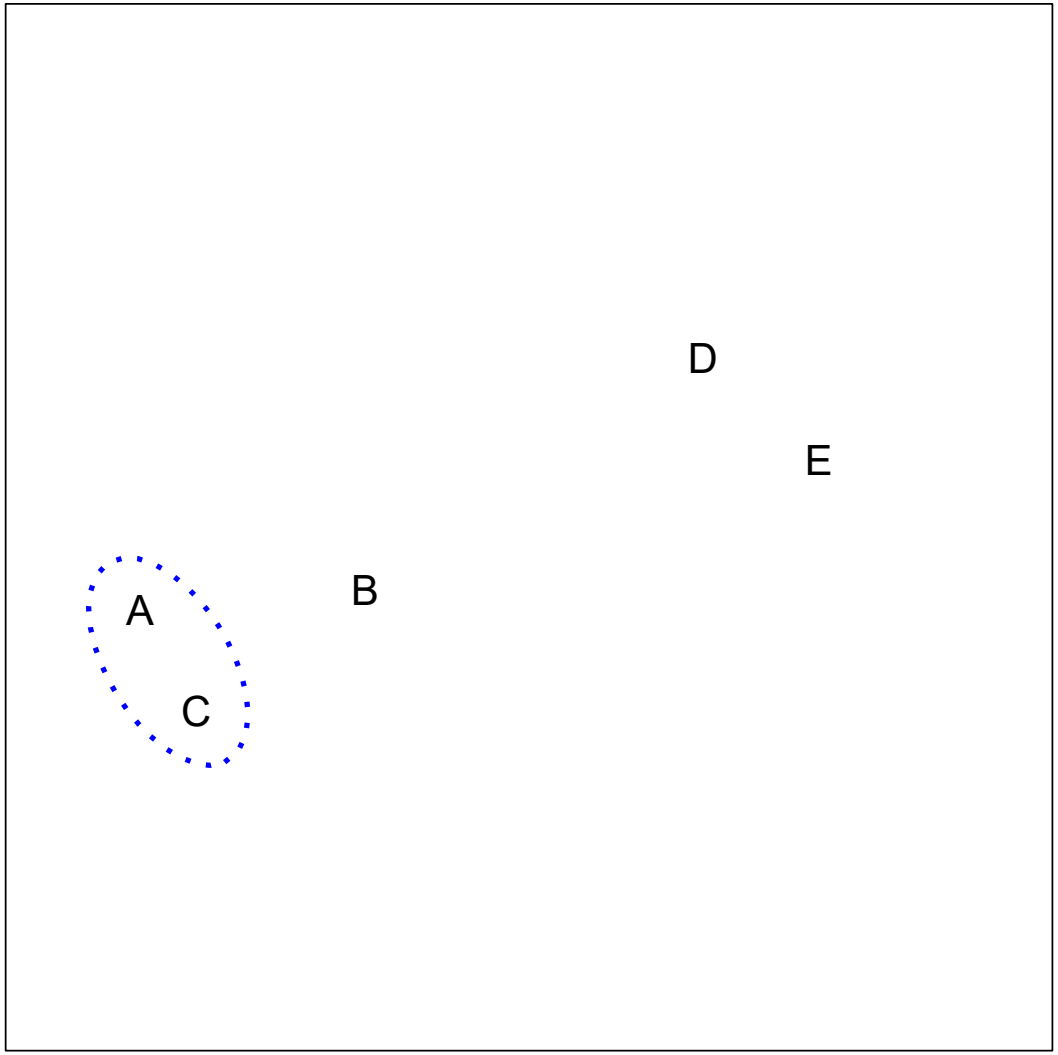
階層的クラスタリング: アイデア

「ボトムアップ」な方法で階層を作る



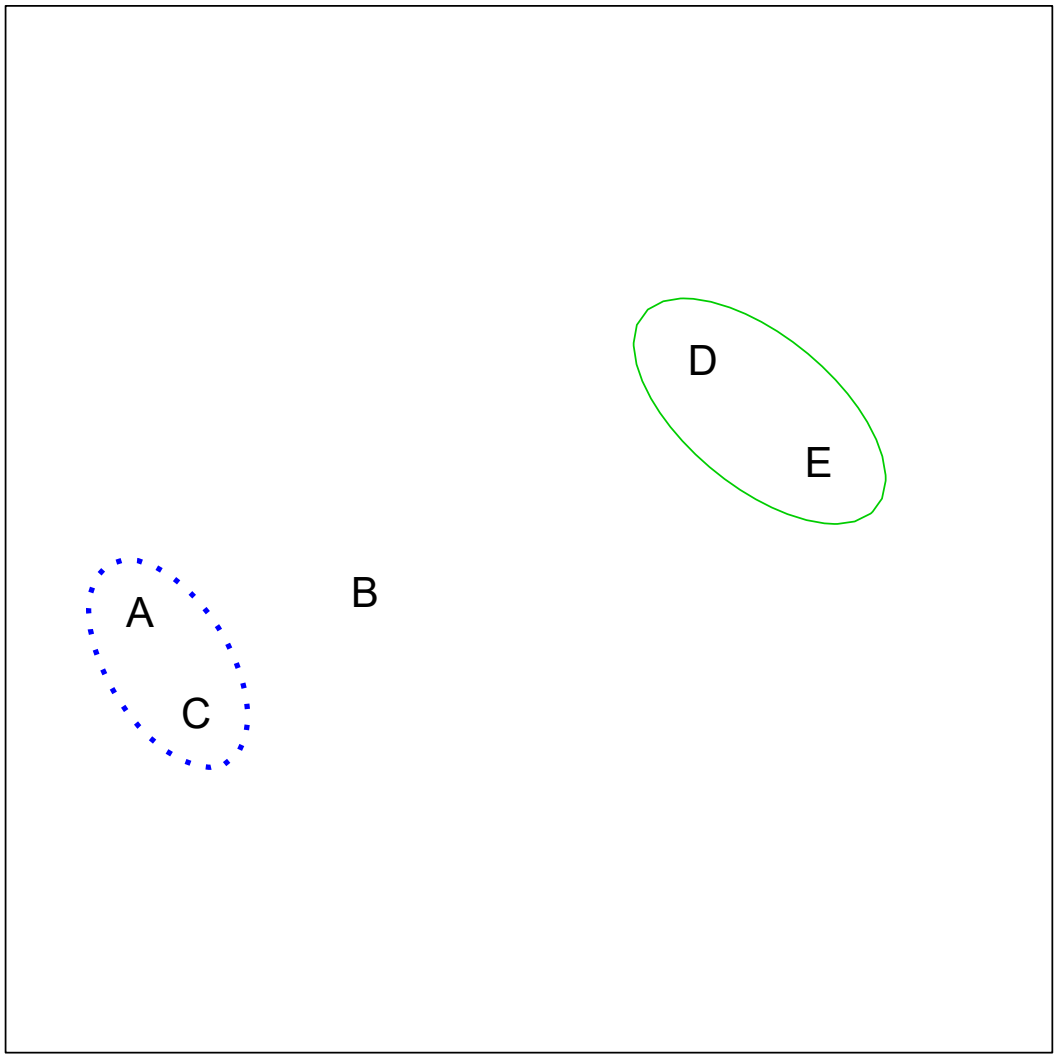
階層的クラスタリング: アイデア

「ボトムアップ」な方法で階層を作る



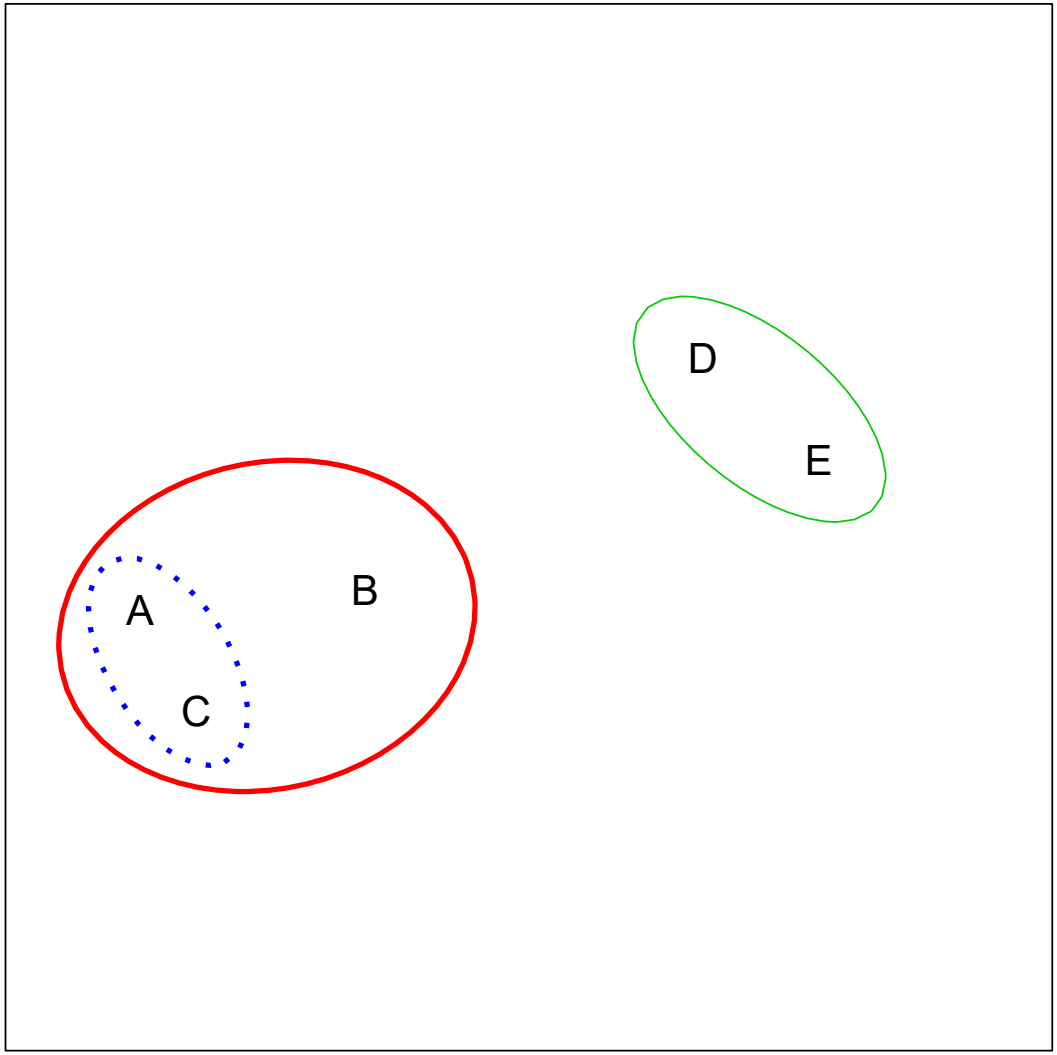
階層的クラスタリング: アイデア

「ボトムアップ」な方法で階層を作る



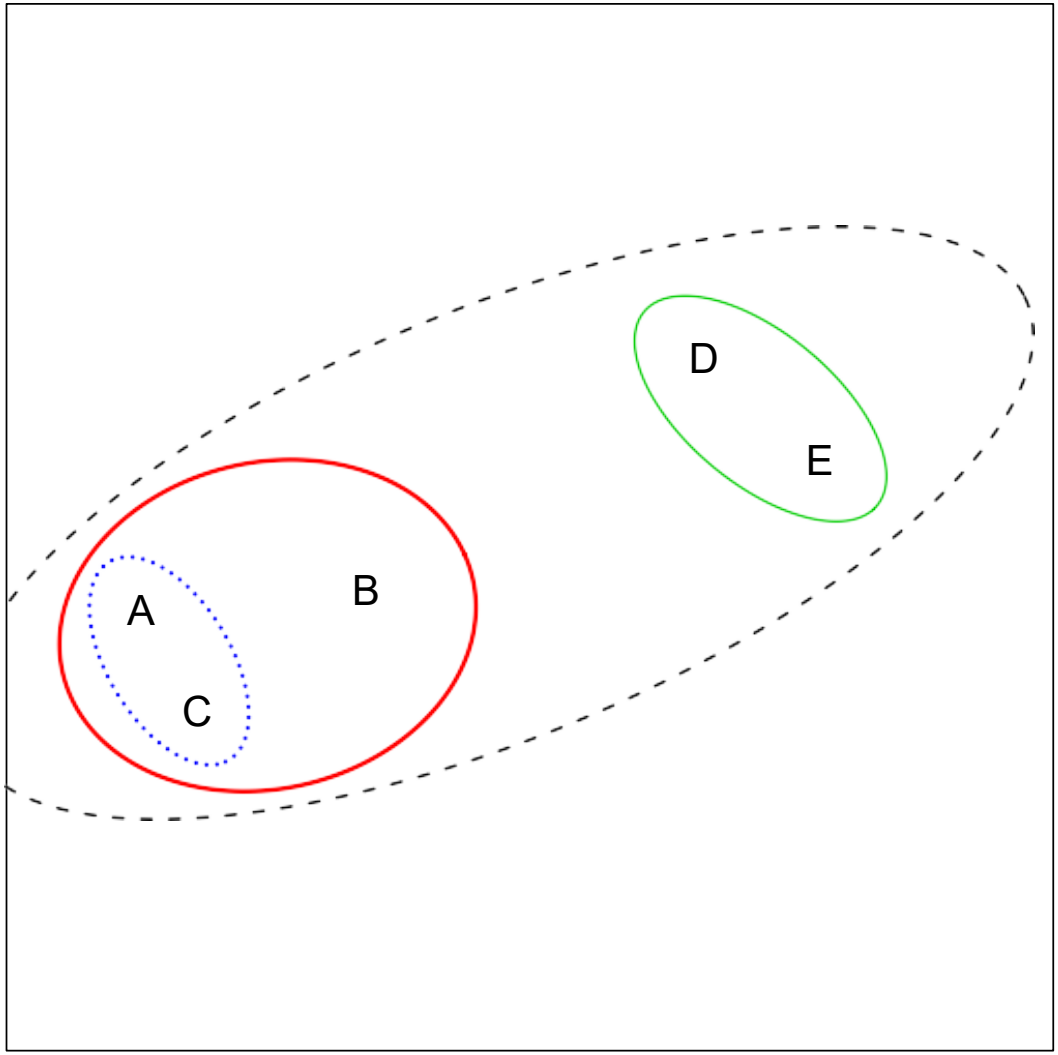
階層的クラスタリング: アイデア

「ボトムアップ」な方法で階層を作る



階層的クラスタリング: アイデア

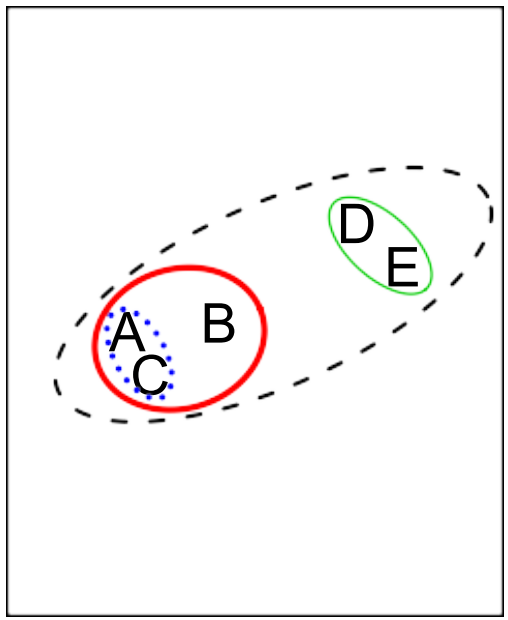
「ボトムアップ」な方法で階層を作る

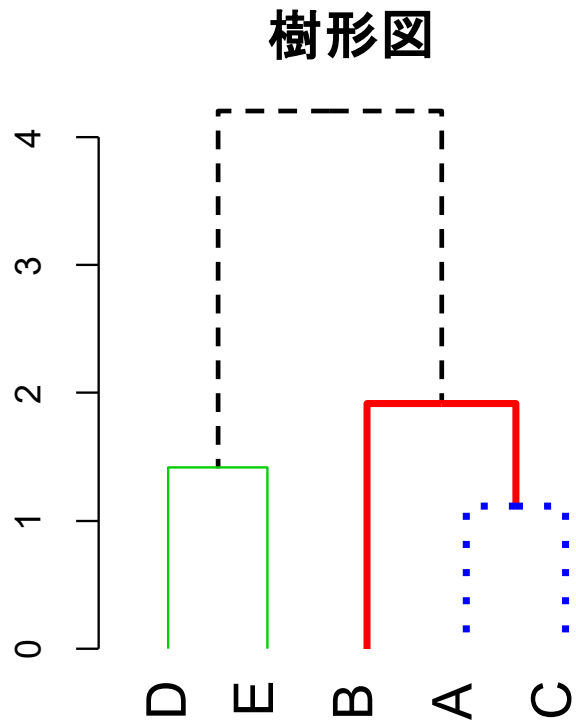


階層的クラスタリング アルゴリズム

言葉では次のように説明:

- 各点自体がクラスタである所から始まる.
- 最も近い2つのクラスタを統合する.
- 繰り返す.
- すべての点が1つのクラスタになったら終わり.

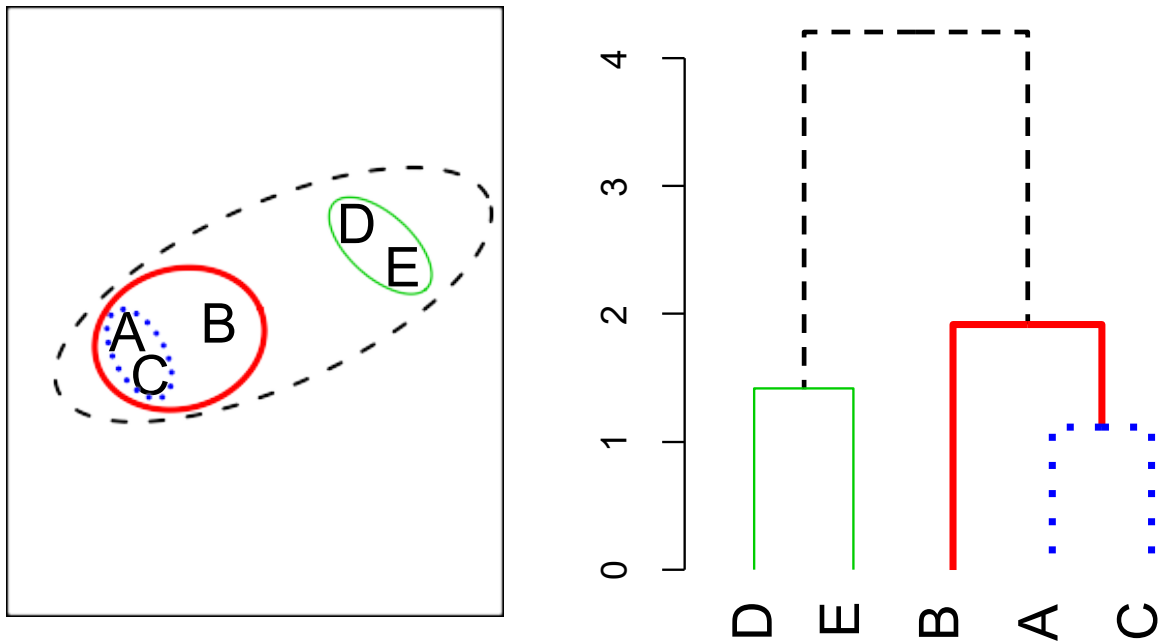




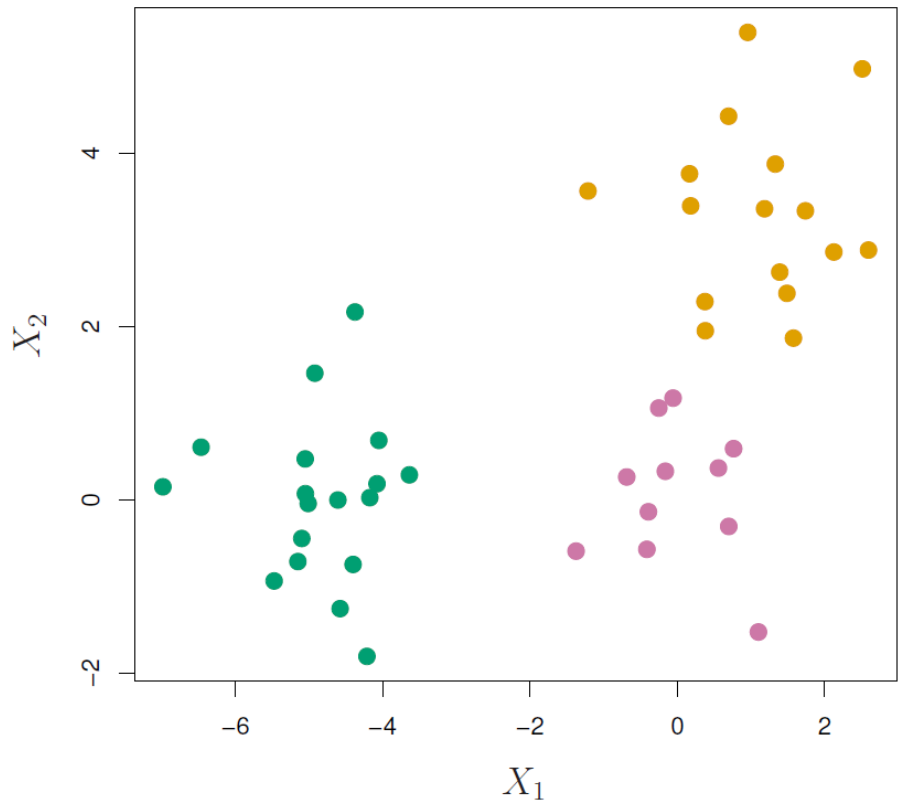
階層的クラスタリング アルゴリズム

言葉では次のように説明:

- 各点自体がクラスタである所から始まる.
- 最も近い2つのクラスタを統合する.
- 繰り返す.
- すべての点が1つのクラスタになったら終わり.

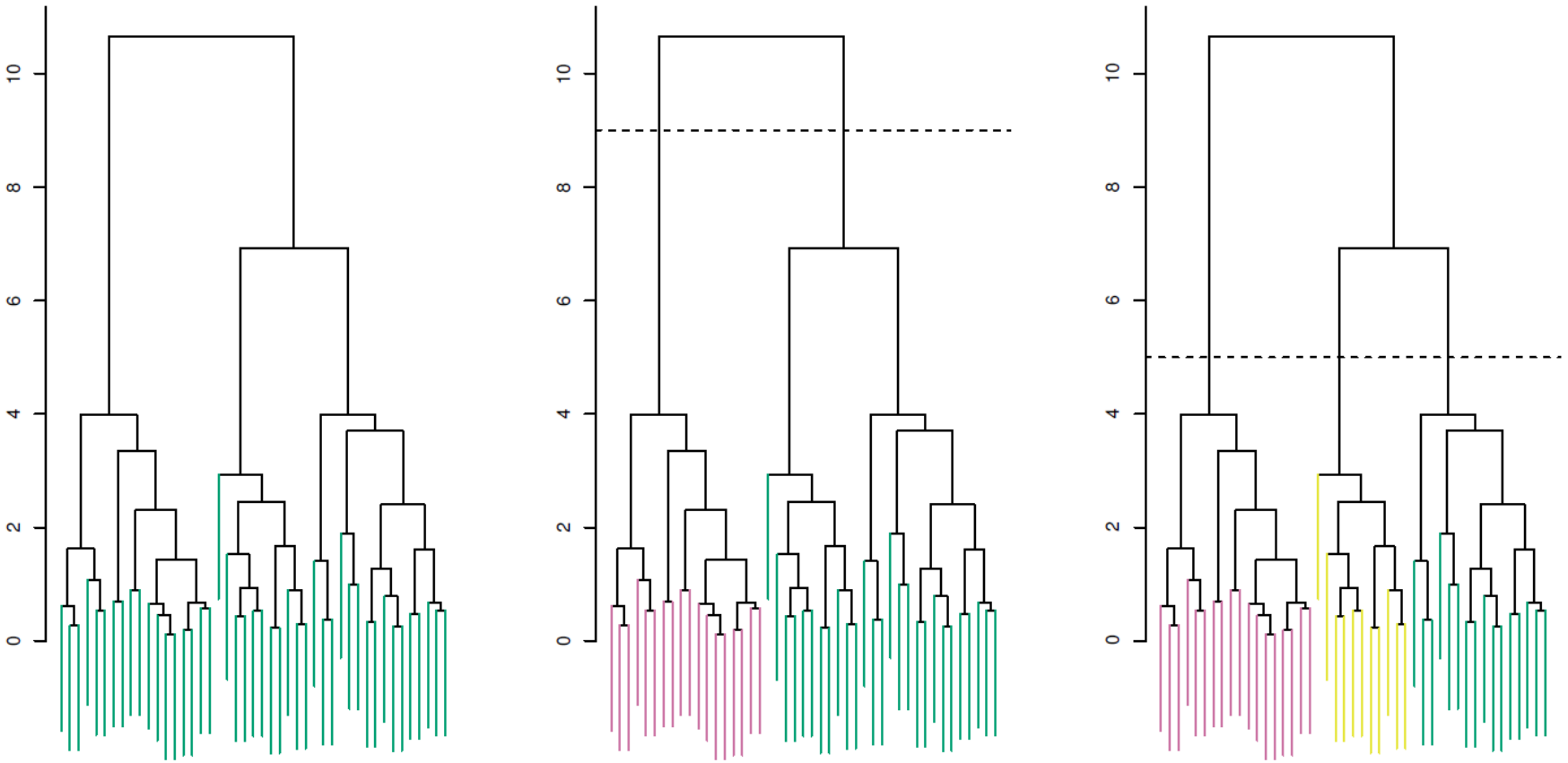


例



2次元空間上の45個の観測. 実際には、色で分けられた3つのクラスがある. しかし、このクラスのラベルを知らないものとして、観測をクラスタリングすることで、データからクラスを発見したい.

階層的クラスタリングの適用



前の図の説明

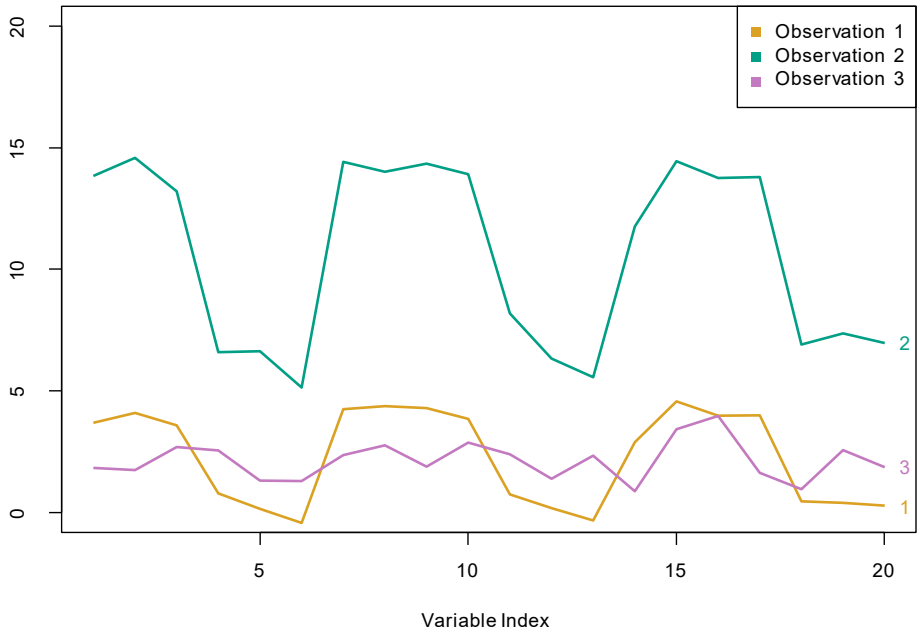
- *左:* 前のスライドのデータを階層的クラスタリングして得られた樹形図. 完全な連結、ユークリッド距離による.
- *真ん中:* 左の樹形図を高さ9で切ったもの (破線で示されている). 異なる色で示される、2つのクラスタを得る.
- *右:* 高さ5で切ったもの. 異なる色で示される、3つのクラスタを得る. 色はクラスタリングに用いておらず、図示のためにのみ用いている.

連結の種類

連結	*説明*
完全	クラスタ同士の最大非類似度．クラスタAの観測とクラスタBの観測の非類似度を全てのペアについて計算する．そして*最大*の非類似度を記録する．
単一	クラスタ同士の最小非類似度．クラスタAの観測とクラスタBの観測の非類似度を全てのペアについて計算する．そして*最小*の非類似度を記録する．
平均	クラスタ同士の平均非類似度．クラスタAの観測とクラスタBの観測の非類似度を全てのペアについて計算する．そして*平均*の非類似度を記録する．
重心	クラスタAの重心(長さpの平均ベクトル)とクラスタBの重心の非類似度．重心連結は、望ましくない*反転*.を起こす事がある．

非類似度の測り方の選択

- 今までユークリッド距離を用いてきた.
- 他の方法は、2つの観測の特徴量が強く相関していれば、似ていると考える、*相関に基づいた距離*を用いる方法.
- これは相関の変わった使用法である. 通常変数の間で計算されるが、ここでは観測のペアに対する観測のプロファイルの間で計算される

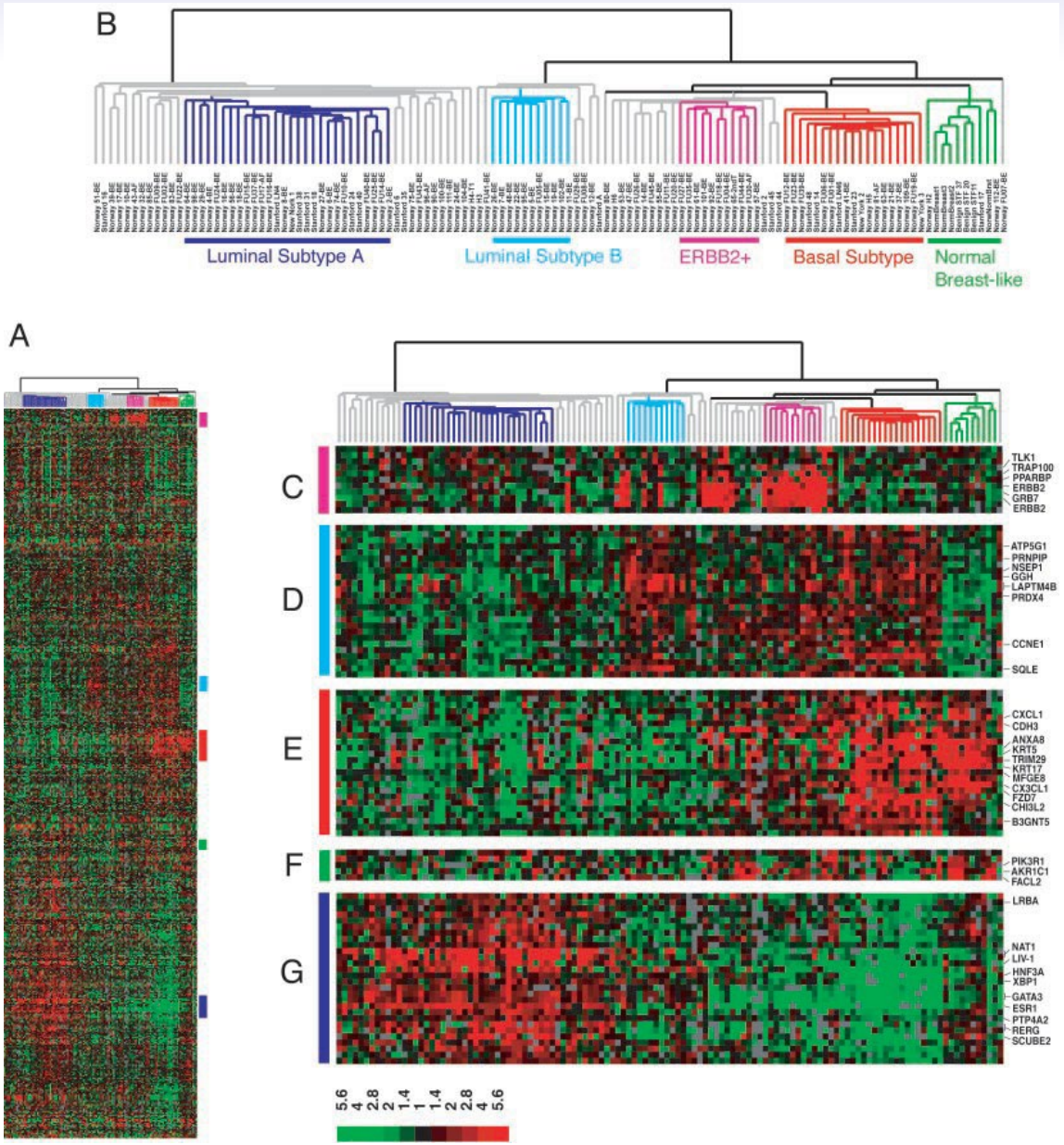


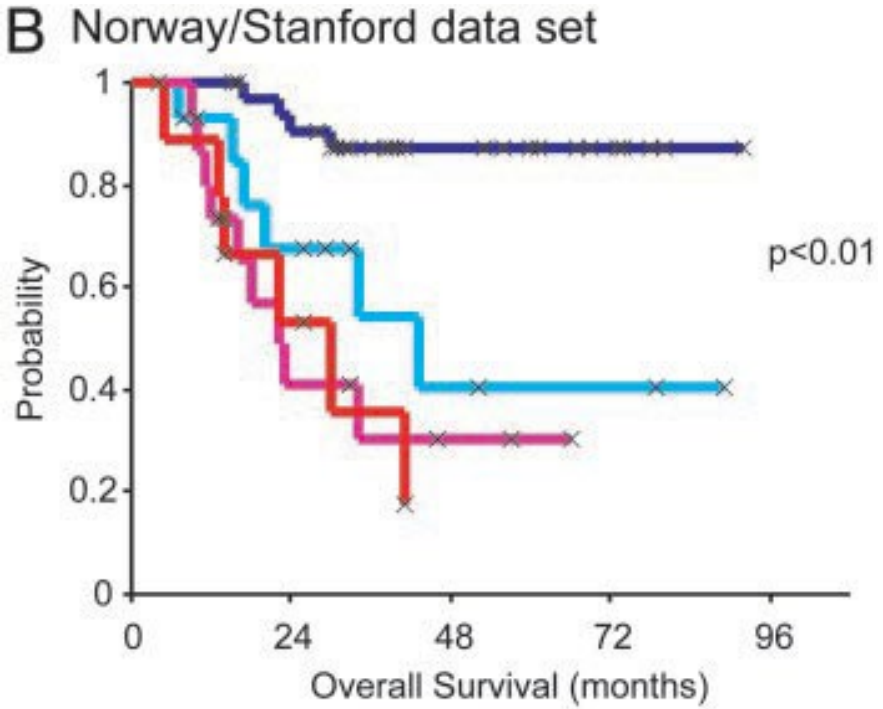
実用上の問題

- *変数のスケールが問題となる!* 観測や特徴は初めにどのように標準化されるべきか？例えば、変数は平均0となるように中心化され、標準偏差1となるようにスケーリングされるべきかもしれない.
- 階層的クラスタリングでは、
  - どのような非類似度が用いられるべきか？
  - どのような連結が用いられるべきか？
- クラスタの数をどのように選ぶか？ (K平均でも階層的クラスタリングでも). 難しい問題である. 誰もが納得する方法はない. より詳しくは、Elements of Statistical Learning, 13章を見よ.
- どの特徴をクラスタリングのために用いるべきか？

例: 乳がんマイクロアレイ研究

- “Repeated observation of breast tumor subtypes in independent gene expression data sets;” Sorlie at el, PNAS 2003
- 88人の乳がん患者の、~約8000の遺伝子に関する遺伝子表現データ.
- 平均連結、相関距離
- 500個の*固有遺伝子*を使ってクラスタリングする. 化学療法の前後で測定を行った. 固有遺伝子は最小の変動を持つ.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GGH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CHI3L2
B3GNT5
PIK3R1
AKR1C1
FACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6


B Norway/Stanford data set
Probability
1
0.8
0.6
0.4
0.2
0
0
24
48
72
96
Overall Survival (months)
p<0.01

結論

- *教師なし学習*は変動の理解やラベルなしデータのグループ化で重要であり、教師あり学習の前処理として用いる事が出来る.
- *教師あり学習* より固有に難しい. それは(結果変数のような)絶対的基準がなく、(テストセットでの精度のような)唯一の目標もない.
- *自己組織写像、独立成分分析、スペクトルクラスタリング*のように、最近多くの発展がある、活発な研究分野である. *The Elements of Statistical Learning*, 14章を見よ.

第13章 : 多重仮説検定
-Multiple Testing-

- 仮説検定をめぐる課題
- 仮説検定の復習
- 多重検定の導入
- Family Wise Error Rate
- Bonferroni補正
- Holmの方法
- 誤発見率
- Benjamini Hochbergの方法
- リサンプリング法

第13章 : 多重仮説検定
-Multiple Testing-

- この章では多重仮説検定を扱う.

多重仮説検定

- この章では多重仮説検定を扱う.
- 単一の帰無仮説は、例えば H_0: コントロール群と処置群のネズミで血圧の期待値が等しい というようなもの.

多重仮説検定

- この章では多重仮説検定を扱う.
- 単一の帰無仮説は、例えば H_0: コントロール群と処置群のネズミで血圧の期待値が等しい というようなもの.
- m個の帰無仮説$H_{01}, \ldots, H_{0m}$を考える. 例えば H_{0j}: コントロール群と処置群のネズミでj番目のバイオマーカーが等しい というようなもの.

多重仮説検定

- この章では多重仮説検定を扱う.
- 単一の帰無仮説は、例えば H_0: コントロール群と処置群のネズミで血圧の期待値が等しい というようなもの.
- m個の帰無仮説$H_{01}, \ldots, H_{0m}$を考える. 例えば H_{0j}: コントロール群と処置群のネズミでj番目のバイオマーカーが等しい というようなもの.
- この設定で，誤って帰無仮説を棄却することがないように気を付ける必要がある. つまり、偽陽性が多くならないようにする.

仮説検定の簡単な復習

仮説検定では「はい」か「いいえ」で答えられる質問を扱う

- 線形回帰において、真の係数 β_j が0に等しいか
- コントロール群と処置群のネズミで血圧の期待値が等しいか

仮説検定の簡単な復習

仮説検定では「はい」か「いいえ」で答えられる質問を扱う

- 線形回帰において、真の係数 β_j が0に等しいか
- コントロール群と処置群のネズミで血圧の期待値が等しいか

仮説検定は以下の手順で行う:

1. 帰無仮説と対立仮説を定める

仮説検定の簡単な復習

仮説検定では「はい」か「いいえ」で答えられる質問を扱う

- 線形回帰において、真の係数 β_j が0に等しいか
- コントロール群と処置群のネズミで血圧の期待値が等しいか

仮説検定は以下の手順で行う:

1. 帰無仮説と対立仮説を定める
2. 検定統計量を定める

仮説検定の簡単な復習

仮説検定では「はい」か「いいえ」で答えられる質問を扱う

- 線形回帰において、真の係数 β_j が0に等しいか
- コントロール群と処置群のネズミで血圧の期待値が等しいか

仮説検定は以下の手順で行う:

1. 帰無仮説と対立仮説を定める
2. 検定統計量を定める
3. p値を計算する

仮説検定の簡単な復習

仮説検定では「はい」か「いいえ」で答えられる質問を扱う

- 線形回帰において、真の係数 β_j が0に等しいか
- コントロール群と処置群のネズミで血圧の期待値が等しいか

仮説検定は以下の手順で行う:

1. 帰無仮説と対立仮説を定める
2. 検定統計量を定める
3. p値を計算する
4. 帰無仮説を棄却するか否かを決定する

1. 帰無仮説と対立仮説を定める

- 帰無仮説 と 対立仮説 に分ける.
- 帰無仮説 H_0 は前もって定める仮説である. 例えば:
  1. 真の係数 β_j が0に等しい.
  2. 血圧の期待値が等しい.

1. 帰無仮説と対立仮説を定める

- 帰無仮説 と 対立仮説 に分ける.
- 帰無仮説 H_0 は前もって定める仮説である. 例えば:
  1. 真の係数 β_j が0に等しい.
  2. 血圧の期待値が等しい.
- 対立仮説 H_a は異なる説や意外な説である. 例えば:
  1. 真の係数 β_j が0でない.
  2. 血圧の期待値が異なる.

2. 検定統計量を定める

- 検定統計量は、手持ちのデータと帰無仮説 H_0 の一致度合いを表す.

2. 検定統計量を定める

- 検定統計量は、手持ちのデータと帰無仮説 H_0 の一致度合いを表す.
- $\hat{\mu}_t / \hat{\mu}_c$ がそれぞれ処置群とコントロール群の n_t / n_c 匹のネズミに対する血圧の標本平均を表す.

2. 検定統計量を定める

- 検定統計量は、手持ちのデータと帰無仮説 H_0 の一致度合いを表す.
- $\hat{\mu}_t / \hat{\mu}_c$ がそれぞれ処置群とコントロール群の n_t / n_c 匹のネズミに対する血圧の標本平均を表す.
- 検定 $H_0 : \mu_t = \mu_c$ のため、2標本 t統計量

$$T = \frac{\hat{\mu}_t - \hat{\mu}_c}{s\sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$$

3. p値を計算する

- p値は H_0が正しいという仮定の下で, 検定統計量が観測値と同等かより極端な値をとる確率である.

3. p値を計算する

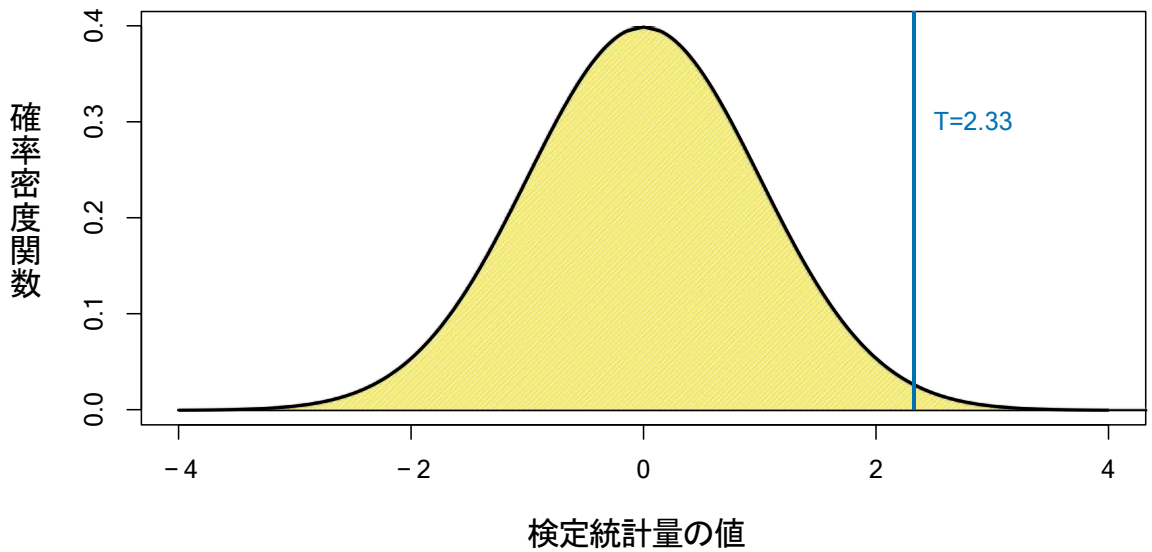
- p値は H_0が正しいという仮定の下で, 検定統計量が観測値と同等かより極端な値をとる確率である.
- 小さな p値はデータが H_0 に反する という根拠を与える.

3. p値を計算する

- p値は H_0が正しいという仮定の下で, 検定統計量が観測値と同等かより極端な値をとる確率である.
- 小さな p値はデータが H_0 に反する という根拠を与える.
- $H_0: \mu_t = \mu_c$ の検定で $T = 2.33$ となったとする.

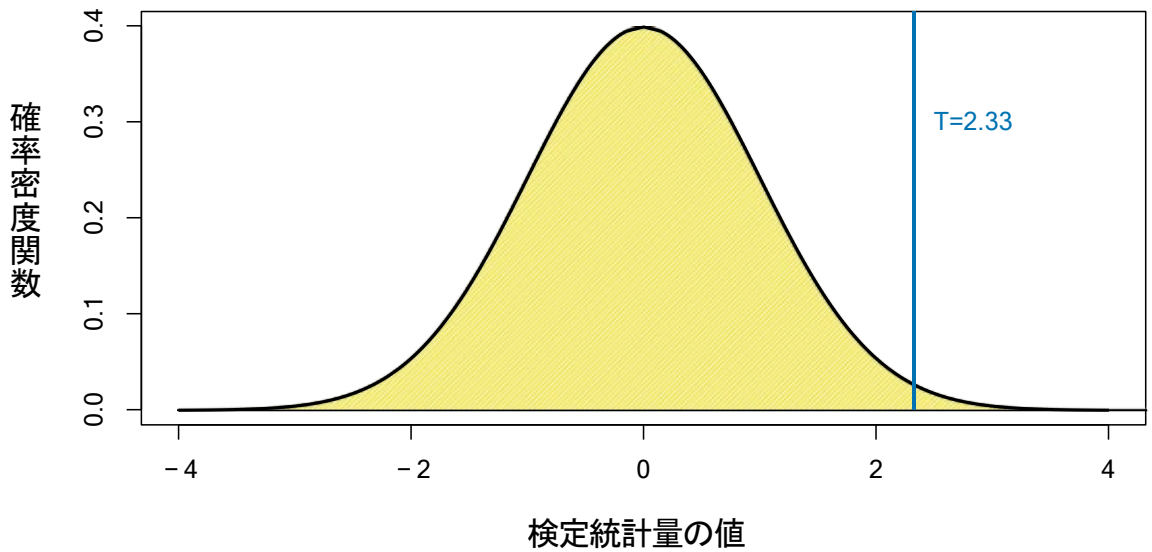
3. p値を計算する

- p値は H_0が正しいという仮定の下で, 検定統計量が観測値と同等かより極端な値をとる確率である.
- 小さな p値はデータが H_0 に反する という根拠を与える.
- $H_0: \mu_t = \mu_c$ の検定で $T = 2.33$ となったとする.
- H_0の下で、2標本t統計量が$T \sim N(0,1)$.



3. p値を計算する

- p値は H_0が正しいという仮定の下で, 検定統計量が観測値と同等かより極端な値をとる確率である.
- 小さな p値はデータが H_0 に反する という根拠を与える.
- $H_0: \mu_t = \mu_c$ の検定で $T = 2.33$ となったとする.
- H_0の下で、2標本t統計量が$T \sim N(0,1)$.



- p値は 0.02 である. なぜなら H_0 が正しいとき$|T|$が2.33より大きくなるのは2%だからである.

4. H_0を棄却するか決定する パート 1

- 小さな p値は H_0の下では検定統計量がそのような大きな値をあまりとらない事を意味する.

4. H_0を棄却するか決定する パート 1

- 小さな p値は H_0の下では検定統計量がそのような大きな値をあまりとらない事を意味する.
- よって、小さな p値はH_0に反する根拠となる.

4. H_0を棄却するか決定する パート 1

- 小さな p値は H_0の下では検定統計量がそのような大きな値をあまりとらない事を意味する.
- よって、小さな p値はH_0に反する根拠となる.
- p値が十分小さいとき, H_0を棄却したい. (そしてその結果、「発見」につながるかもしれない).

4. H_0を棄却するか決定する パート 1

- 小さな p値は H_0の下では検定統計量がそのような大きな値をあまりとらない事を意味する.
- よって、小さな p値はH_0に反する根拠となる.
- p値が十分小さいとき, H_0を棄却したい. (そしてその結果、「発見」につながるかもしれない).
- しかしどれだけ小さければ良いか. これに答えるためには, 第１種の過誤について理解する必要がある.

4. H_0を棄却するか決定する パート2

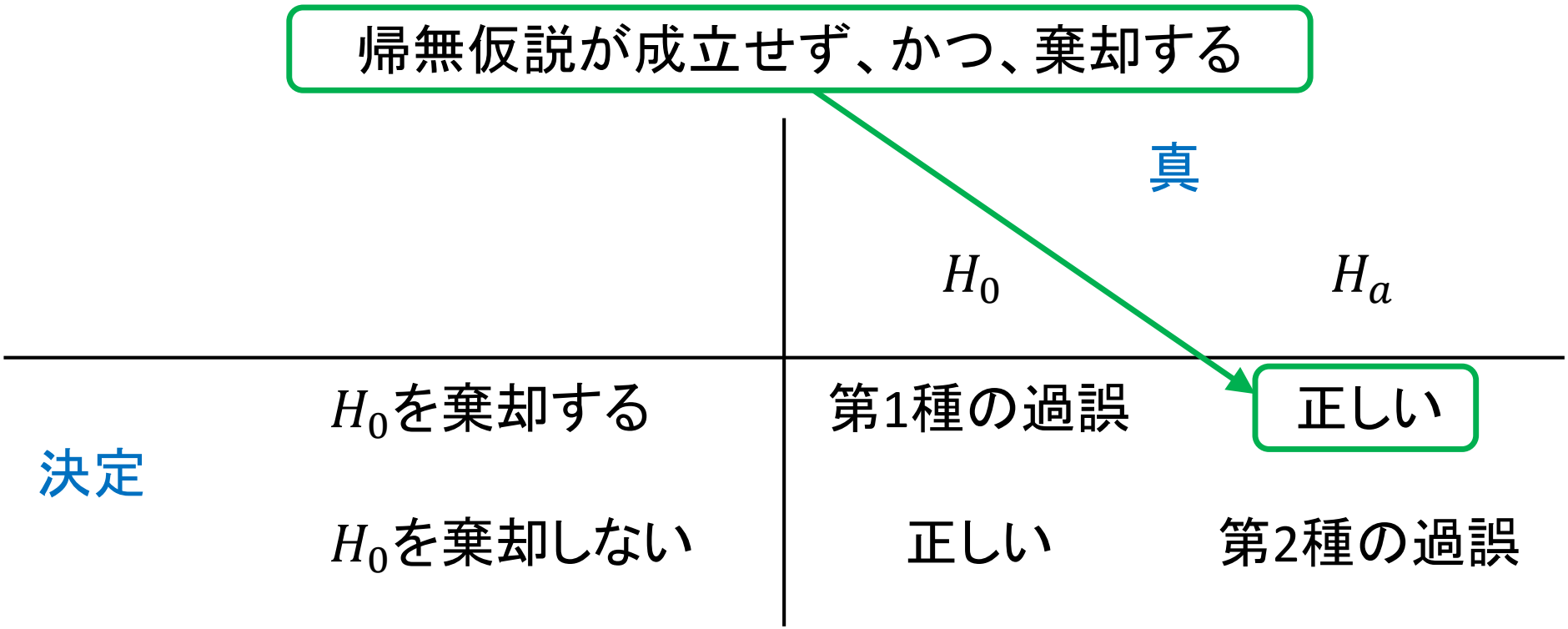
		真	
		H_0	H_a
決定	H_0を棄却する	第1種の過誤	正しい
	H_0を棄却しない	正しい	第2種の過誤

4. H_0を棄却するか決定する パート2

帰無仮説が成立し、かつ、棄却しない

		真	
		H_0	H_a
決定	H_0を棄却する	第1種の過誤	正しい
	H_0を棄却しない	正しい	第2種の過誤

4. H_0を棄却するか決定する パート 2



4. H_0を棄却するか決定する パート2

帰無仮説が成立せず、かつ、棄却しない

		真	
		H_0	H_a
決定	H_0を棄却する	第1種の過誤	正しい
	H_0を棄却しない	正しい	第2種の過誤

4. H_0を棄却するか決定する パート2

帰無仮説が成立し、かつ、棄却する

		真	
		H_0	H_a
決定	H_0を棄却する	第1種の過誤	正しい
	H_0を棄却しない	正しい	第2種の過誤

4. H_0を棄却するか決定する パート 3

- 第一種の過誤率 は第一種の過誤を起こす確率である.
- 第一種の過誤率を小さくしたい.

4. H_0を棄却するか決定する パート 3

- 第一種の過誤率 は第一種の過誤を起こす確率である.
- 第一種の過誤率を小さくしたい.
- p値が αより小さいときにのみH_0を棄却するとする. このとき, 第一種の過誤率は高々αである.

4. H_0を棄却するか決定する パート 3

- 第一種の過誤率 は第一種の過誤を起こす確率である.
- 第一種の過誤率を小さくしたい.
- p値が αより小さいときにのみH_0を棄却するとする. このとき, 第一種の過誤率は高々αである.
- よって, p値がαより小さいときに H_0を棄却する: αは $0.05, 0.01, 0.001$とすることが多い.

多重検定

- m個の帰無仮説$H_{01}, \ldots, H_{0m}$を検定したいとする.

多重検定

- m個の帰無仮説$H_{01}, \ldots, H_{0m}$を検定したいとする.
- 単にそれぞれの帰無仮説に対するp値が (例えば) 0.01より小さいときに棄却すれば良いか.

多重検定

- m個の帰無仮説$H_{01}, \ldots, H_{0m}$を検定したいとする.
- 単にそれぞれの帰無仮説に対するp値が (例えば) 0.01より小さいときに棄却すれば良いか.
- それぞれの帰無仮説に対するp値が0.01より小さいときに棄却すると、第一種の過誤はどれほどになるか.

思考実験

- 歪みのないコインを10回投げたとする. H_0 : コインは歪みがない を検定したい.

思考実験

- 歪みのないコインを10回投げたとする．H_0：コインは歪みがない を検定したい．
  - おそらく表と裏が同じような回数出るだろう．
  - p値は小さくならないだろう．H_0を棄却しない．

思考実験

- 歪みのないコインを10回投げたとする．H_0：コインは歪みがない を検定したい．
  - おそらく表と裏が同じような回数出るだろう．
  - p値は小さくならないだろう．H_0を棄却しない．
- 1024個の歪みのないコインをそれぞれ10回投げるとどうか．

思考実験

- 歪みのないコインを10回投げたとする. H_0 : コインは歪みがない を検定したい.
  - おそらく表と裏が同じような回数出るだろう.
  - p値は小さくならないだろう. H_0を棄却しない.
- 1024個の歪みのないコインをそれぞれ10回投げるとどうか.
  - おそらく1つのコインはすべて裏が出るだろう.

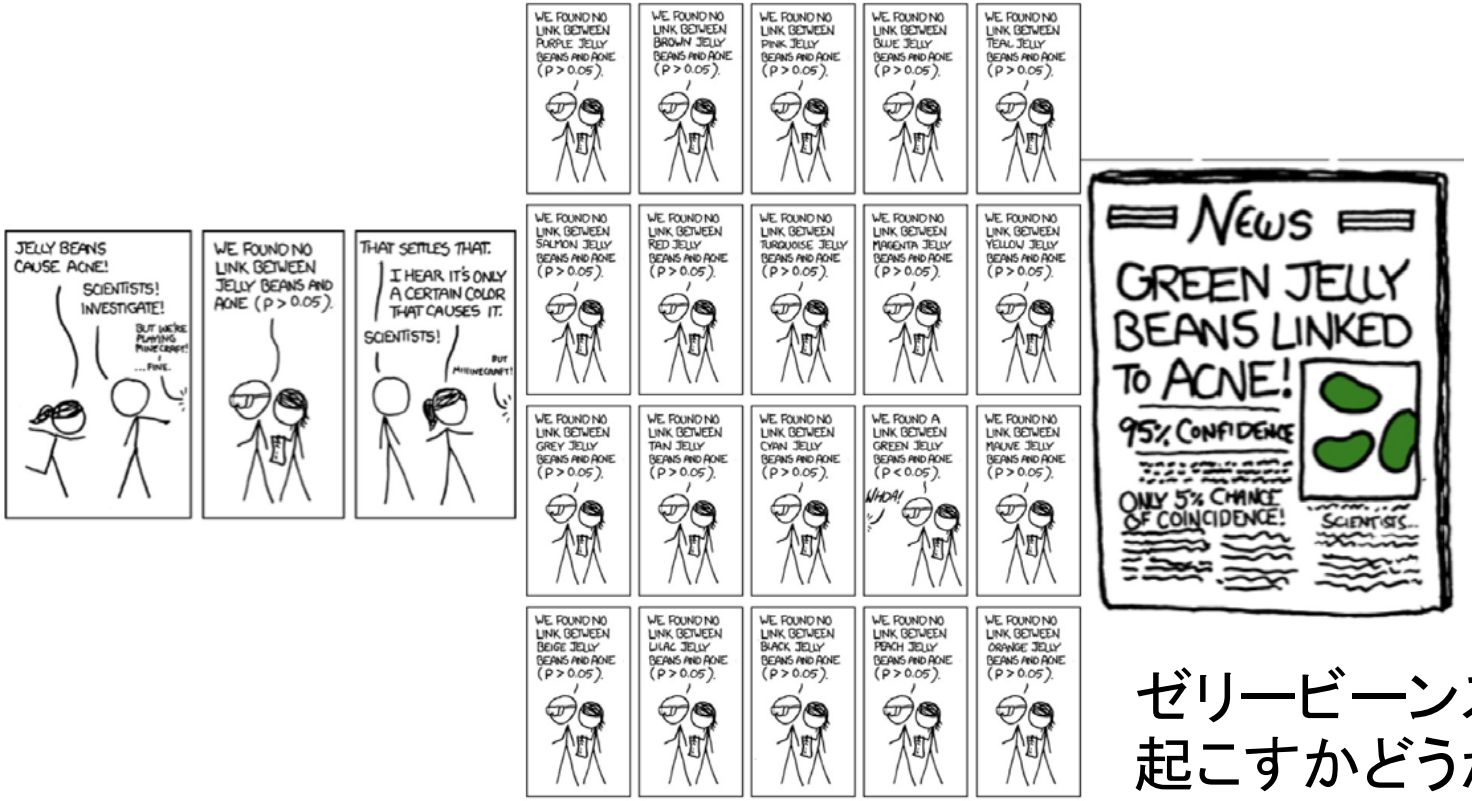
思考実験

- 歪みのないコインを10回投げたとする. H_0 : コインは歪みがない を検定したい.
  - おそらく表と裏が同じような回数出るだろう.
  - p値は小さくならないだろう. H_0 を棄却しない.
- 1024個の歪みのないコインをそれぞれ10回投げるとどうか.
  - おそらく1つのコインはすべて裏が出るだろう.
  - そのコインに対するp値は0.002となる!
  - よって歪みがあると結論付ける, つまり, 歪みがないとしても, H_0 を棄却してしまう.

思考実験

- 歪みのないコインを10回投げたとする. H_0 : コインは歪みがない を検定したい.
  - おそらく表と裏が同じような回数出るだろう.
  - p値は小さくならないだろう. H_0を棄却しない.
- 1024個の歪みのないコインをそれぞれ10回投げるとどうか.
  - おそらく1つのコインはすべて裏が出るだろう.
  - そのコインに対するp値は0.002となる!
  - よって歪みがあると結論付ける, つまり, 歪みがないとしても, H_0 を棄却してしまう.
- 多くの検定を行うと, 偶然によって, たいていとても小さなp値が得られてしまう.

多重検定: ウェブコミックサイトXKCDでの風刺



https://xkcd.com/882/

ゼリービーンズがニキビを引き起こすかどうかを調べる.
20色のゼリービーンズで実験した結果、緑色のゼリービーンズとニキビの間には有意水準5%で関連が見られた.
まさに多重性の問題.

多重検定の困難さ

- (実は正しい)$H_{01}, \ldots, H_{0m}$を検定したいとする. p値が0.01以下である帰無仮説を棄却する.

多重検定の困難さ

- (実は正しい)$H_{01}, \ldots, H_{0m}$を検定したいとする. p値が0.01以下である帰無仮説を棄却する.
- このとき, おおよそ$0.01 \times m$個の帰無仮説は誤って棄却してしまう.

多重検定の困難さ

- (実は正しい)$H_{01}, \ldots, H_{0m}$を検定したいとする. p値が0.01以下である帰無仮説を棄却する.
- このとき, おおよそ$0.01 \times m$個の帰無仮説は誤って棄却してしまう.
- $m = 10{,}000$であれば, 偶然によって, 約100個の帰無仮説は誤って棄却してしまう!

多重検定の困難さ

- (実は正しい)$H_{01}, \ldots, H_{0m}$を検定したいとする. p値が0.01以下である帰無仮説を棄却する.
- このとき, おおよそ$0.01 \times m$個の帰無仮説は誤って棄却してしまう.
- $m = 10{,}000$であれば, 偶然によって, 約100個の帰無仮説は誤って棄却してしまう!
- このようにして, 多くの第一種の過誤, つまり, 偽陽性が生じる.

FWER (Family-Wise Error Rate)

- FWER(family-wise error rate) は m個の仮説検定を行うときに, 少なくとも1つの 第一種の過誤を起こす確率である.

FWER (Family-Wise Error Rate)

- FWER(family-wise error rate) はm個の仮説検定を行うときに, 少なくとも1つの 第一種の過誤を起こす確率である.
- $\mathrm{FWER} = \Pr(V \geq 1)$

	H_0が正しい	H_0が正しくない	合計
H_0を棄却する	V	S	R
H_0を棄却しない	U	W	$m-R$
合計	m_0	$m-m_0$	m

FWERをコントロールすることの困難さ

$$\begin{aligned}\mathrm{FWER} &= 1 - \Pr(\text{どの帰無仮説も誤って棄却しない}) \\ &= 1 - \Pr\left(\cap_{j=1}^{m} \left\{H_{0j}\text{を誤って棄却しない}\right\}\right)\end{aligned}$$

FWERをコントロールすることの困難さ

$$\begin{aligned}\mathrm{FWER} &= 1 - \Pr(\text{どの帰無仮説も誤って棄却しない}) \\ &= 1 - \Pr\left(\cap_{j=1}^{m} \left\{H_{0j}\text{を誤って棄却しない}\right\}\right)\end{aligned}$$

検定が独立で, すべてのH_{0j}が正しいとき,

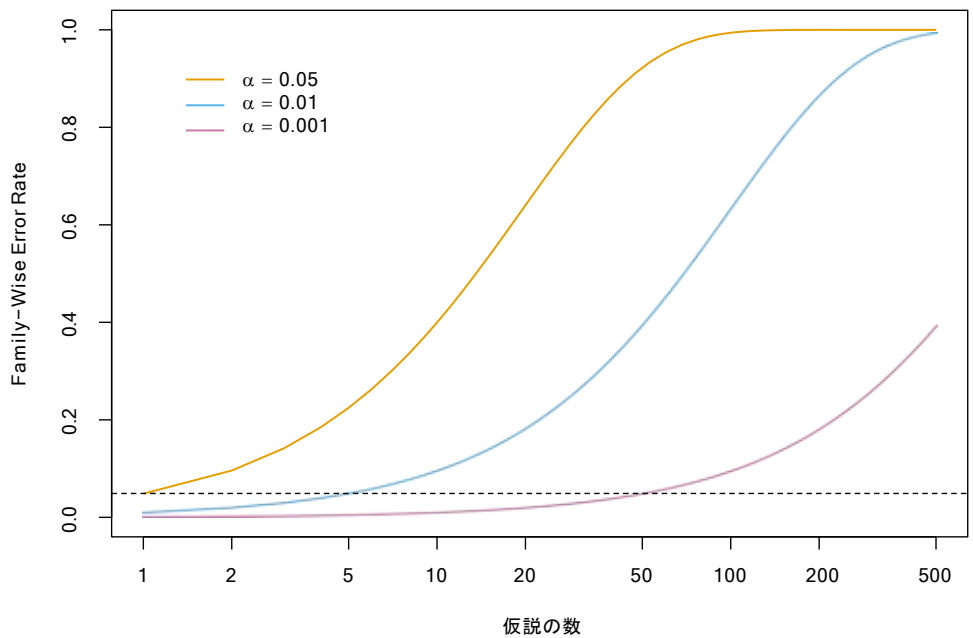
$$\mathrm{FWER} = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m$$

FWERをコントロールすることの困難さ

$$\begin{aligned}\mathrm{FWER} &= 1 - \Pr(\text{どの帰無仮説も誤って棄却しない}) \\ &= 1 - \Pr\left(\cap_{j=1}^{m} \left\{H_{0j}\text{を誤って棄却しない}\right\}\right)\end{aligned}$$

検定が独立で, すべてのH_{0j}が正しいとき,

$$\mathrm{FWER} = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m$$



The Bonferroni補正

$$\begin{aligned}\text{FWER} &= \Pr(\text{少なくとも1つの帰無仮説を誤って棄却する}) \\ &= \Pr\left(\cup_{j=1}^{m} A_j\right) \\ &\leq \sum_{j=1}^{m} \Pr(A_j)\end{aligned}$$

ただし A_j はj番目の帰無仮説を誤って棄却する事象.

The Bonferroni補正

$$\begin{aligned}\text{FWER} &= \Pr(\text{少なくとも1つの帰無仮説を誤って棄却する}) \\ &= \Pr\left(\cup_{j=1}^{m} A_j\right) \\ &\leq \sum_{j=1}^{m} \Pr(A_j)\end{aligned}$$

ただし A_j はj番目の帰無仮説を誤って棄却する事象.

- p値がα/mより小さいときのみ棄却すると,

  $$\text{FWER} \leq \sum_{j=1}^{m} \Pr(A_j) \leq \sum_{j=1}^{m} \frac{\alpha}{m} = m \times \frac{\alpha}{m} = \alpha$$

  なぜなら, $\Pr(A_j) \leq \alpha/m$
- これを Bonferroni補正という. 水準αでFWERをコントロールするために, 各帰無仮説をp値がα/m以下のときに棄却する.

ファンドマネージャーデータ

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

ファンドマネージャーデータ

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

- H_{0j}: j人目のマネージャーの期待超過収益が0である.
- p値が$\alpha = 0.05$のときにH_{0j}を棄却すると, 1人目 と 3人目 のマネージャーは明らかに0でない超過収益を出していると結論づけることになるだろう.

ファンドマネージャーデータ

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

- H_{0j}: j人目のマネージャーの期待超過収益が0である.
- p値が$\alpha = 0.05$のときにH_{0j}を棄却すると, 1人目 と 3人目 のマネージャーは明らかに0でない超過収益を出していると結論づけることになるだろう.
- しかし, 多重検定の結果, FWER は0.05より大きくなる.

ファンドマネージャーデータに対するBonferroni補正

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

- Bonferroni補正により, p値が$\alpha/m = 0.05/5 = 0.01$より小さいとき棄却する.

ファンドマネージャーデータに対するBonferroni補正

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

- Bonferroni補正により, p値が$\alpha/m = 0.05/5 = 0.01$より小さいとき棄却する.
- 結果, 1人目のマネージャーのみ帰無仮説を棄却することになる.

ファンドマネージャーデータに対するBonferroni補正

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

- Bonferroni補正により, p値が$\alpha/m = 0.05/5 = 0.01$より小さいとき棄却する.
- 結果, 1人目のマネージャーのみ帰無仮説を棄却することになる.
- このとき, FWERは高々0.05である.

FWERのコントロールに関するHolmの方法

FWERのコントロールに関するHolmの方法

1. p値 $p_1, \ldots, p_m$ をm個の帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.

FWERのコントロールに関するHolmの方法

1. p値 $p_1, \ldots, p_m$ をm個の帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
2. m個のp値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.

FWERのコントロールに関するHolmの方法

1. p値 $p_1, \ldots, p_m$ をm個の帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
2. m個のp値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.
3. Lを次のように定める

$$L = \min\left\{j : p_{(j)} > \frac{\alpha}{m+1-j}\right\}.$$

FWERのコントロールに関するHolmの方法

1. p値 $p_1, \ldots, p_m$ をm個の帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
2. m個のp値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.
3. Lを次のように定める

$$L = \min\left\{j : p_{(j)} > \frac{\alpha}{m+1-j}\right\}.$$

4. 各仮説H_{0j}を$p_j < p_{(L)}$のときに棄却する.

FWERのコントロールに関するHolmの方法

1. p値 $p_1, \ldots, p_m$ をm個の帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
2. m個のp値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.
3. Lを次のように定める

$$L = \min\left\{j : p_{(j)} > \frac{\alpha}{m+1-j}\right\}.$$

4. 各仮説H_{0j}を$p_j < p_{(L)}$のときに棄却する.

- Holmの方法は水準αでFWERをコントロールする.

ファンドマネージャーデータに対するHolmの方法

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

- p値を並べ替えると, $p_{(1)} = 0.006, p_{(2)} = 0.012,$ $p_{(3)} = 0.601, p_{(4)} = 0.756, p_{(5)} = 0.918.$

ファンドマネージャーデータに対するHolmの方法

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

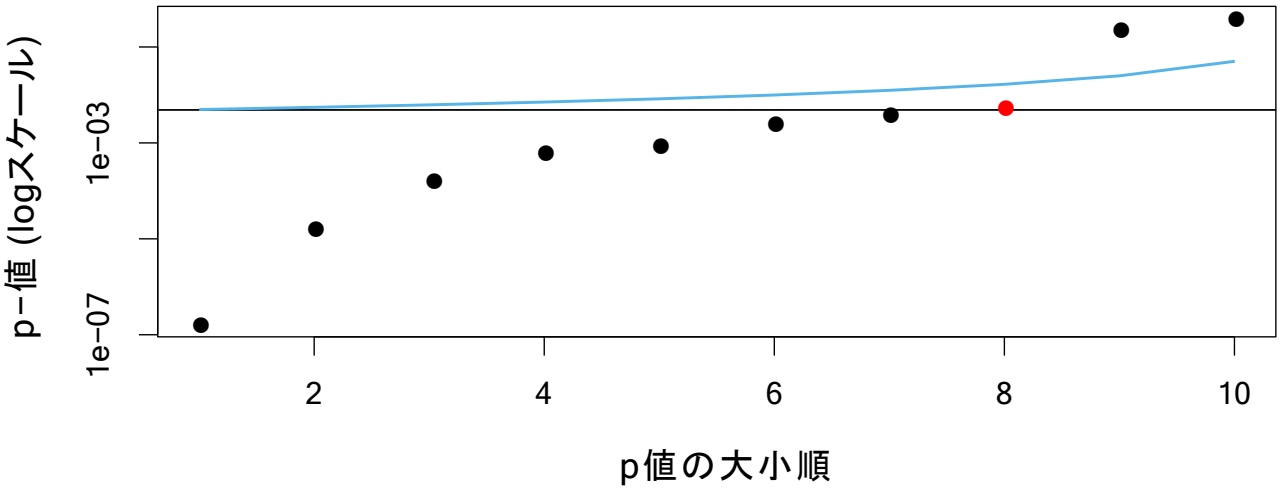
- p値を並べ替えると, $p_{(1)} = 0.006, p_{(2)} = 0.012,$ $p_{(3)} = 0.601, p_{(4)} = 0.756, p_{(5)} = 0.918.$
- Holmの方法では最初の2人の帰無仮説を棄却する. なぜなら,
  - $p_{(1)} = 0.006 < 0.05/(5 + 1 - 1) = 0.0100,$
  - $p_{(2)} = 0.012 < 0.05/(5 + 1 - 2) = 0.0125,$
  - $p_{(3)} = 0.601 < 0.05/(5 + 1 - 3) = 0.0167.$

ファンドマネージャーデータに対するHolmの方法

マネージャー	平均 $\bar{x}$	s	t統計量	p値
1	3.0	7.4	2.86	0.006
2	-0.1	6.9	-0.10	0.918
3	2.8	7.5	2.62	0.012
4	0.5	6.7	0.53	0.601
5	0.3	6.8	0.31	0.756

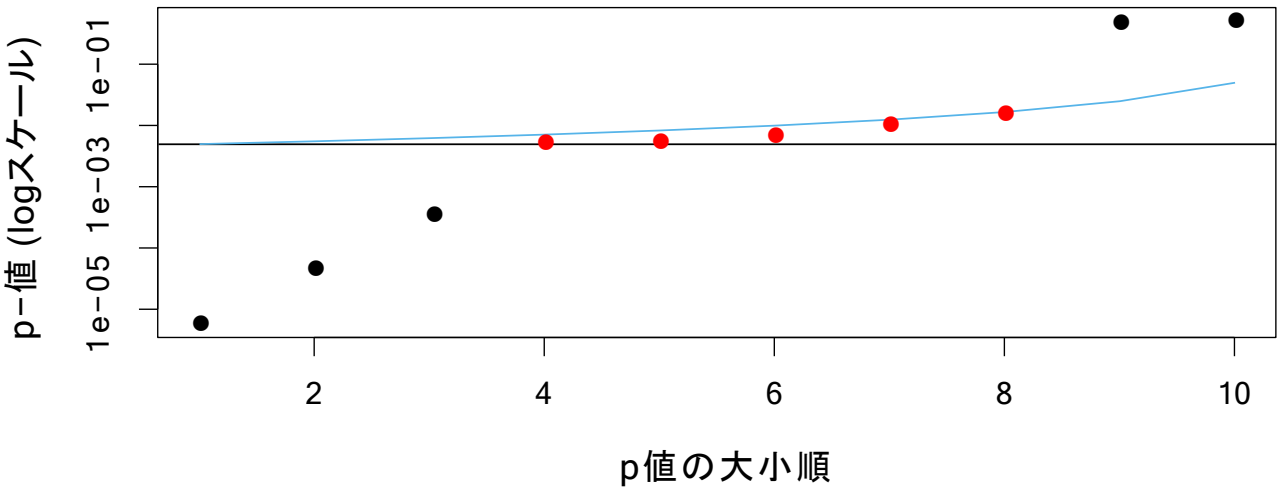
- p値を並べ替えると, $p_{(1)} = 0.006, p_{(2)} = 0.012,$ $p_{(3)} = 0.601, p_{(4)} = 0.756, p_{(5)} = 0.918.$
- Holmの方法では最初の2人の帰無仮説を棄却する. なぜなら,
  - $p_{(1)} = 0.006 < 0.05/(5+1-1) = 0.0100,$
  - $p_{(2)} = 0.012 < 0.05/(5+1-2) = 0.0125,$
  - $p_{(3)} = 0.601 < 0.05/(5+1-3) = 0.0167.$
- Holmの方法では1人目と3人目のマネージャーに対してH_0を棄却するが, Bonferroniの方法では1人目のマネージャーに対してのみH_0を棄却する.

$m = 10$でのp値の比較



- FWERを0.05でコントロールする.
- 黒線より下のp値は, Bonferroni補正で棄却される.
- 青線より下のp値は, Holmの方法で棄却される.
- Holmの方法とBonferroniの方法は, 黒点では同じ結論となるが, Holmの方法のみ赤点を棄却する.

より極端な例



- 5つの仮説がHolmの方法では棄却されるが, Bonferroni補正では棄却されない.
- ともに, FWERを0.05でコントロールしているのだが.

Holmの方法かBonferroni補正か

- Bonferroni補正はシンプルで，p値がα/mより小さい仮説はすべて棄却する．
- Holmの方法は少し複雑だが，FWERをコントロールしつつもより多くの棄却を導く．
- よって，Holmの方法の方が良い選択である．

他の方法

- FWER をコントロールする方法は仮説の状況に応じ多数ある.

他の方法

- FWER をコントロールする方法は仮説の状況に応じ多数ある.
- 例えば:
  - Tukeyの方法: 群の間での母平均の差の対比較に対する方法.

他の方法

- FWER をコントロールする方法は仮説の状況に応じ多数ある.
- 例えば:
  - Tukeyの方法: 群の間での母平均の差の対比較に対する方法.
  - Schefféの方法: 母平均の線形結合に対する検定, 例えば,

$$H_0: \frac{1}{2}(\mu_1 + \mu_3) = \frac{1}{3}(\mu_2 + \mu_4 + \mu_5)$$

  に対する方法.

他の方法

- FWERをコントロールする方法は仮説の状況に応じ多数ある.
- 例えば:
  - Tukeyの方法: 群の間での母平均の差の対比較に対する方法.
  - Schefféの方法: 母平均の線形結合に対する検定, 例えば,

$$H_0: \frac{1}{2}(\mu_1 + \mu_3) = \frac{1}{3}(\mu_2 + \mu_4 + \mu_5)$$

    に対する方法.

- Bonferroni補正やHolmの方法は一般的方法で多くの状況で機能するが, 特殊な条件下では, Tukeyの方法やSchefféの方法のような方法がより良い結果を与えうる:
  つまり, FWERをコントロールしつつより多くの棄却を生むことがある.

誤発見率

誤発見率

- 次の表に戻って考える:

	H_0が正しい	H_0が正しくない	合計
H_0を棄却する	V	S	R
H_0を棄却しない	U	W	$m-R$
合計	m_0	$m-m_0$	m

誤発見率

- 次の表に戻って考える:

	H_0が正しい	H_0が正しくない	合計
H_0を棄却する	V	S	R
H_0を棄却しない	U	W	$m-R$
合計	m_0	$m-m_0$	m

- FWER は $\Pr(V>1)$, つまり, いずれかの 帰無仮説を誤って棄却してしまう確率 に注目している.

誤発見率

- 次の表に戻って考える:

	H_0が正しい	H_0が正しくない	合計
H_0を棄却する	V	S	R
H_0を棄却しない	U	W	$m-R$
合計	m_0	$m-m_0$	m

- FWERは$\Pr(V \geq 1)$, つまり, いずれかの 帰無仮説を誤って棄却してしまう確率 に注目している.
- mが大きいとき, これは厳しい要請である. とても保守的(つまり, たまにしか棄却しないこと)になってしまう.

誤発見率

- 次の表に戻って考える:

	H_0が正しい	H_0が正しくない	合計
H_0を棄却する	V	S	R
H_0を棄却しない	U	W	$m-R$
合計	m_0	$m-m_0$	m

- FWERは$\Pr(V \geq 1)$, つまり, いずれかの 帰無仮説を誤って棄却してしまう確率 に注目している.
- mが大きいとき, これは厳しい要請である. とても保守的(つまり, たまにしか棄却しないこと)になってしまう.
- 代わりに, 誤発見率(FDR, False Discovery Rate)をコントロールすることもある.

$$\mathrm{FDR} = \mathrm{E}(V/R).$$

誤発見率の背後にある直観

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = E\left(\frac{\text{誤って棄却した数}}{\text{棄却した数}}\right)$$

誤発見率の背後にある直観

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = E\left(\frac{\text{誤って棄却した数}}{\text{棄却した数}}\right)$$

- ある科学者が$m = 20,000$の薬の候補に対し, それぞれ仮説検定を行う.

誤発見率の背後にある直観

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = E\left(\frac{\text{誤って棄却した数}}{\text{棄却した数}}\right)$$

- ある科学者が$m = 20,000$の薬の候補に対し, それぞれ仮説検定を行う.
- この科学者は発見のために有用であろう候補の集合を特定したい.

誤発見率の背後にある直観

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = E\left(\frac{\text{誤って棄却した数}}{\text{棄却した数}}\right)$$

- ある科学者が$m = 20,000$の薬の候補に対し，それぞれ仮説検定を行う．
- この科学者は発見のために有用であろう候補の集合を特定したい．
- 本当に「有用である」，つまり，H_0を誤って棄却している事の少ない集合となっているという安心があれば良い．

誤発見率の背後にある直観

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = E\left(\frac{\text{誤って棄却した数}}{\text{棄却した数}}\right)$$

- ある科学者が$m = 20,000$の薬の候補に対し，それぞれ仮説検定を行う.
- この科学者は発見のために有用であろう候補の集合を特定したい.
- 本当に「有用である」，つまり，H_0を誤って棄却している事の少ない集合となっているという安心があれば良い.
- FWERは$\Pr$(少なくとも一つ誤って棄却してしまう)をコントロールする.

誤発見率の背後にある直観

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\right) = E\left(\frac{\text{誤って棄却した数}}{\text{棄却した数}}\right)$$

- ある科学者が$m = 20,000$の薬の候補に対し, それぞれ仮説検定を行う.
- この科学者は発見のために有用であろう候補の集合を特定したい.
- 本当に「有用である」, つまり, H_0を誤って棄却している事の少ない集合となっているという安心があれば良い.
- FWERはPr(少なくとも一つ誤って棄却してしまう)をコントロールする.
- FDRは, 本当は間違って棄却してしまった候補の割合をコントロールする. これがこの科学者の欲しいものである.

FDRをコントロールするBenjamini-Hochbergの方法

FDRをコントロールするBenjamini-Hochbergの方法

1. FDR をコントロールする水準qを定める.

FDRをコントロールするBenjamini-Hochbergの方法

1. FDRをコントロールする水準qを定める.
2. p値$p_1, \ldots, p_m$を各帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.

FDRをコントロールするBenjamini-Hochbergの方法

1. FDRをコントロールする水準qを定める.
2. p値$p_1, \ldots, p_m$を各帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
3. p値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.

FDRをコントロールするBenjamini-Hochbergの方法

1. FDRをコントロールする水準qを定める.
2. p値$p_1, \ldots, p_m$を各帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
3. p値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.
4. Lを$L = \max\{j : p_{(j)} < qj/m\}$と定める.

FDRをコントロールするBenjamini-Hochbergの方法

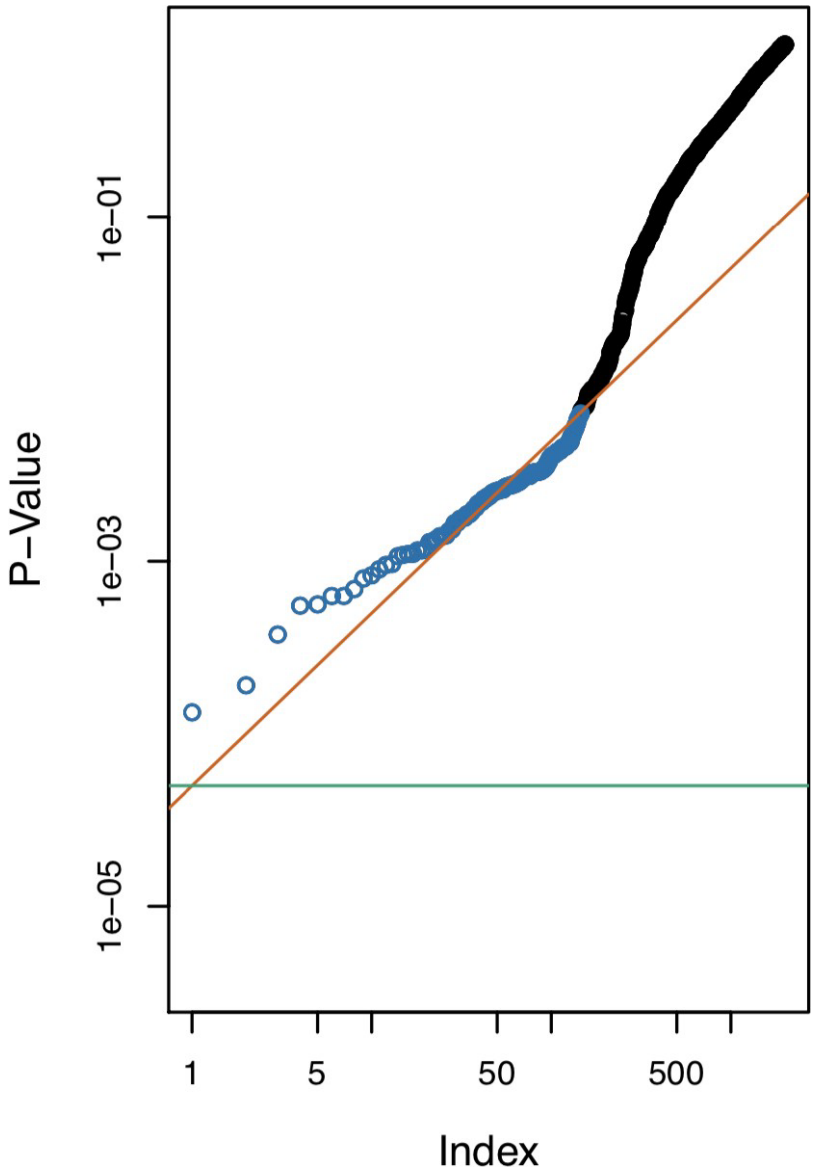
1. FDRをコントロールする水準qを定める.
2. p値$p_1, \ldots, p_m$を各帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
3. p値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.
4. Lを$L = \max\{j : p_{(j)} < qj/m\}$と定める.
5. $p_j \leq p_{(L)}$となる帰無仮説H_{0j}をすべて棄却する.

FDRをコントロールするBenjamini-Hochbergの方法

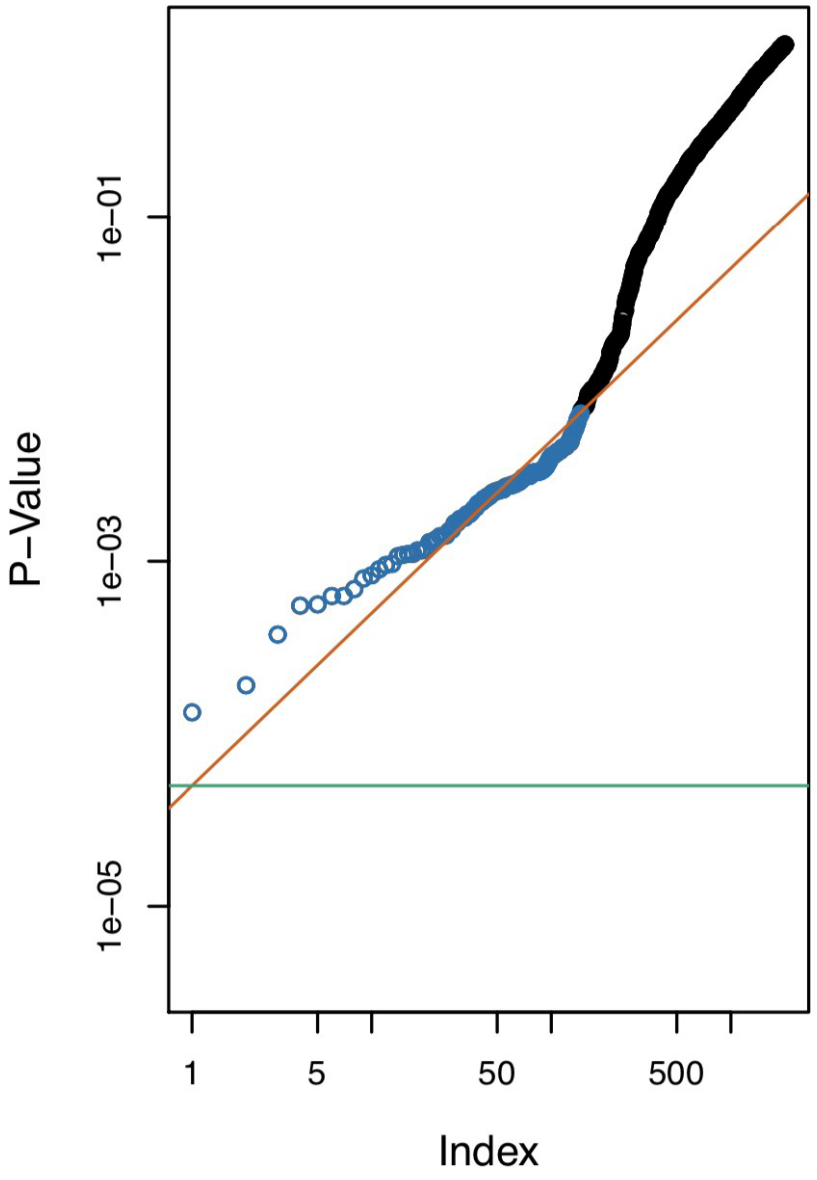
1. FDRをコントロールする水準qを定める.
2. p値$p_1, \ldots, p_m$を各帰無仮説$H_{01}, \ldots, H_{0m}$に対して計算する.
3. p値を$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$と並べる.
4. Lを$L = \max\{j : p_{(j)} < qj/m\}$と定める.
5. $p_j \leq p_{(L)}$となる帰無仮説H_{0j}をすべて棄却する.

このとき, $\mathrm{FDR} \leq q$.

FDRとFWERの 比 較 パート 1

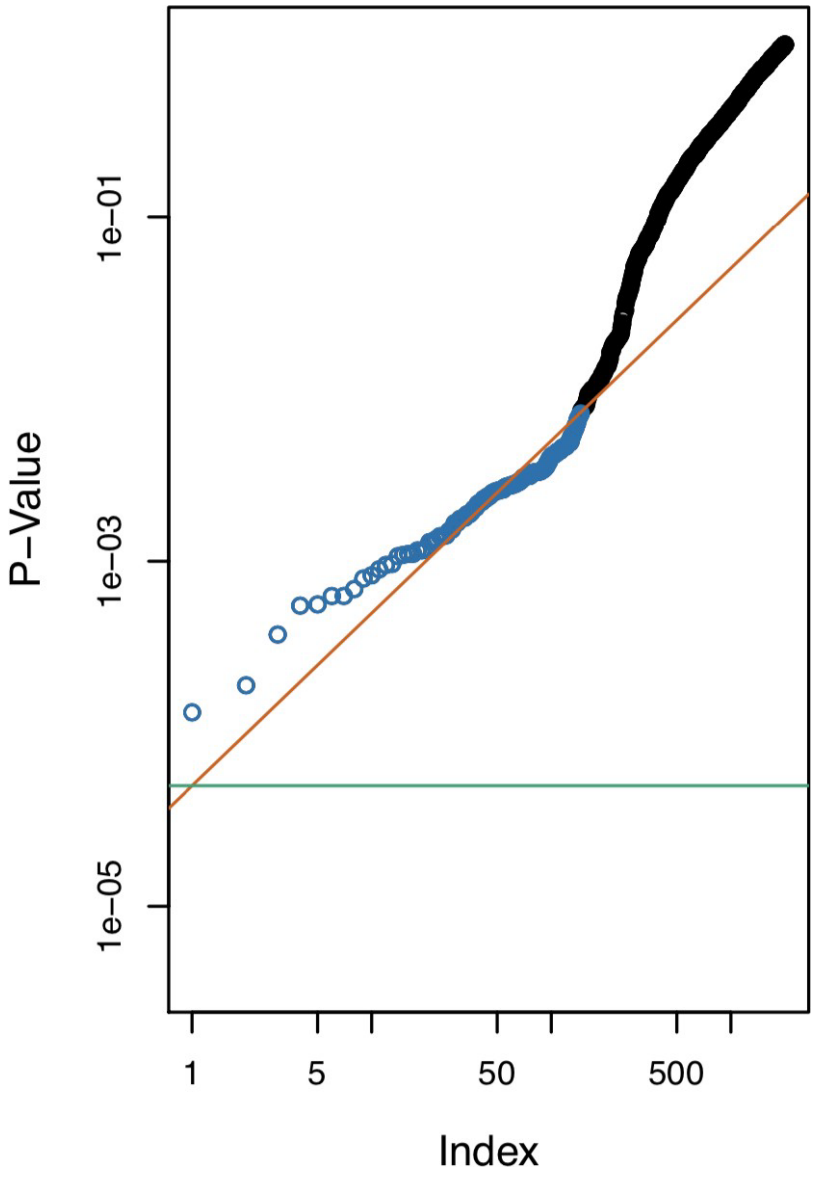


FDRとFWERの比較 パート 1



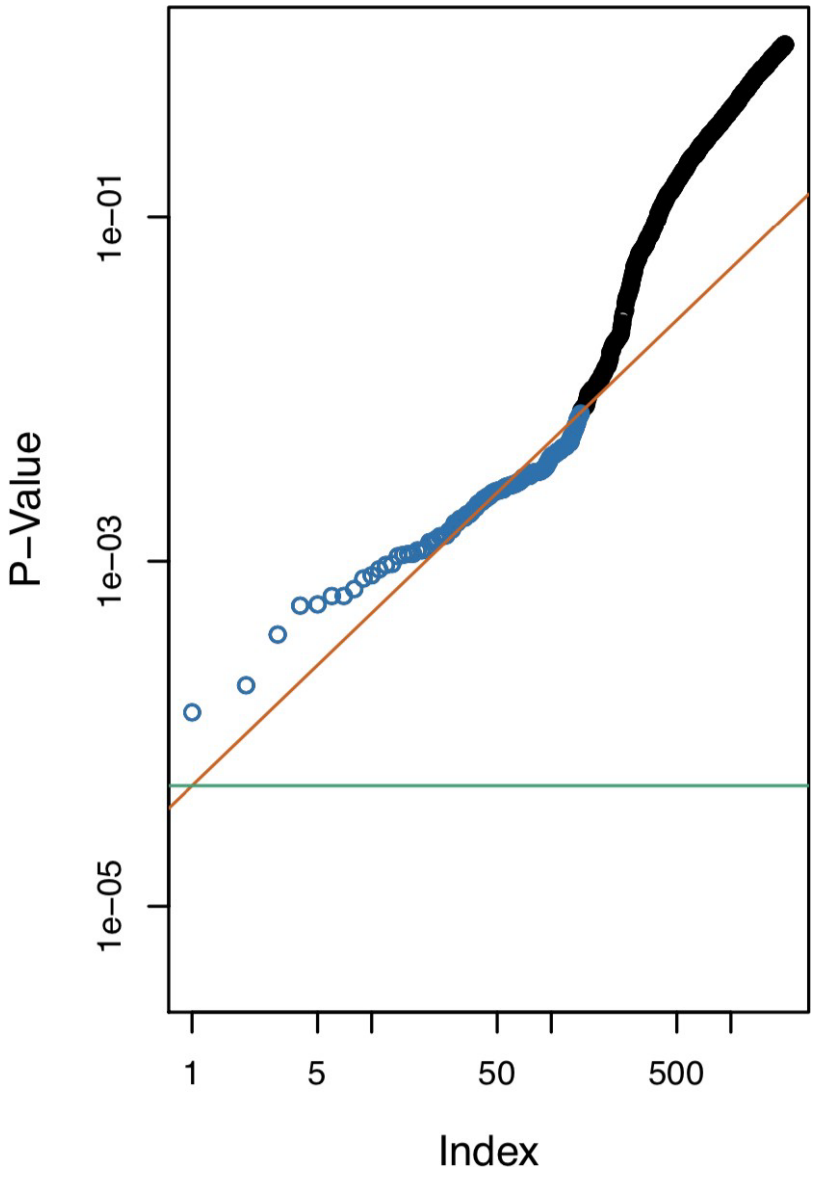
- $m = 2,000$の帰無仮説に対するp値を図示している.

FDRとFWERの比較 パート 1



- $m = 2,000$の帰無仮説に対するp値を図示している.
- Bonferroni補正により, 水準$\alpha = 0.1$でFWERをコントロールする: 緑線より下だと棄却する. (棄却されない!)

FDRとFWERの比較 パート 1



- $m = 2,000$の帰無仮説に対するp値を図示している.
- Bonferroni補正により, 水準$\alpha = 0.1$でFWERをコントロールする: 緑線より下だと棄却する. (棄却されない!)
- 水準$q = 0.1$でBenjamini-Hochbergの方法によりFDRをコントロールする: 青点の仮説は棄却される.

FDRとFWERの比較 パート 2

FDRとFWERの 比 較 パート 2

- ファンドデータによる$m = 5$のp値は以下の通り:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$

FDRとFWERの 比 較 パート 2

- ファンドデータによる$m = 5$のp値は以下の通り:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$
- このとき, $p_{(1)} = 0.006, p_{(2)} = 0.012, p_{(3)} = 0.601,$
  $p_{(4)} = 0.756,\ p_{(5)} = 0.918.$

FDRとFWERの 比 較 パート 2

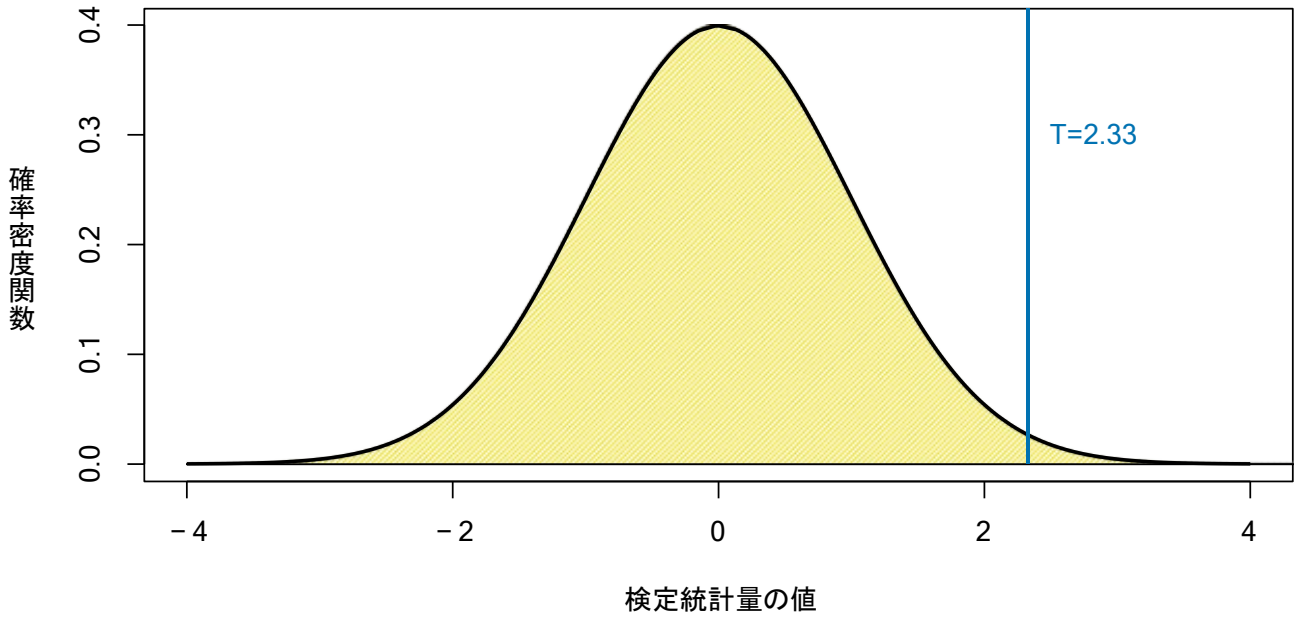
- ファンドデータによる$m = 5$のp値は以下の通り:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$
- このとき, $p_{(1)} = 0.006, p_{(2)} = 0.012, p_{(3)} = 0.601,$
  $p_{(4)} = 0.756,\ p_{(5)} = 0.918.$
- 水準$\mathrm{q} = 0.05$でBenjamini-Hochbergの方法に
  よりFDRをコントロールする:
  - $p_{(1)} < 0.05/5, p_{(2)} < 2 \times 0.05/5, p_{(3)} > 3 \times 0.05/5,$
    $p_{(4)} > 4 \times 0.05/5, p_{(5)} > 5 \times 0.05/5$である.
  - H_{01}, H_{03}を棄却する.

FDRとFWERの比較 パート 2

- ファンドデータによる$m = 5$のp値は以下の通り:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$
- このとき, $p_{(1)} = 0.006, p_{(2)} = 0.012, p_{(3)} = 0.601,$
  $p_{(4)} = 0.756,\ p_{(5)} = 0.918.$
- 水準$\mathrm{q} = 0.05$でBenjamini-Hochbergの方法によりFDRをコントロールする:
  - $p_{(1)} < 0.05/5, p_{(2)} < 2 \times 0.05/5, p_{(3)} > 3 \times 0.05/5,$
    $p_{(4)} > 4 \times 0.05/5, p_{(5)} > 5 \times 0.05/5$である.
  - H_{01}, H_{03}を棄却する.
- 水準$\alpha = 0.05$でBonferroni補正を用いてFWERをコントロールする:
  - p値が$0.05/5$より小さい帰無仮説は棄却される.
  - H_{01}のみ棄却する.

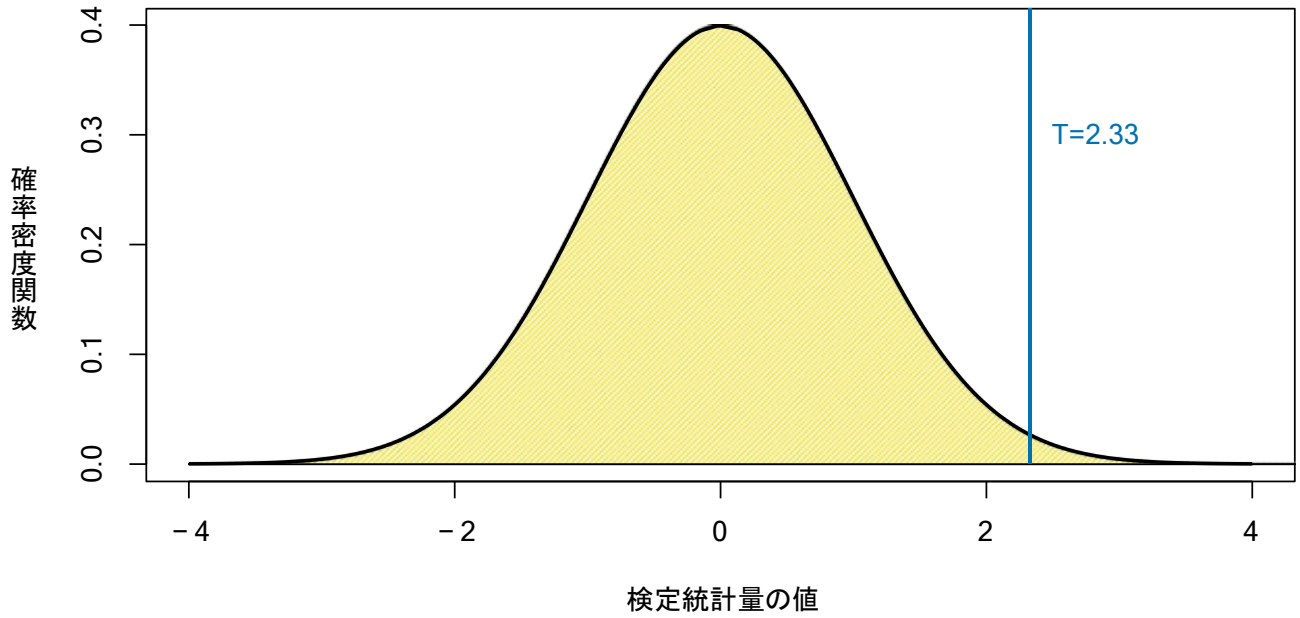
リサンプリングによる方法

- ここまで, 帰無仮説H_0に対して検定統計量Tを用いて検定を行う際, H_0の下での統計量Tの分布が分かっている状況を考えてきた.
- 分布を分かっていることからp値を計算できた.



リサンプリングによる方法

- ここまで, 帰無仮説H_0に対して検定統計量Tを用いて検定を行う際, H_0の下での統計量Tの分布が分かっている状況を考えてきた.
- 分布を分かっていることからp値を計算できた.



- 理論的な帰無分布が分からない場合はどうか.

2標本t検定に対するリサンプリングによる方法, パート 1

- $H_0: E(X) = E(Y)$と$H_a: E(X) \neq E(Y)$の検定を, Xからのn_X個の独立な観測とYからのn_Y個の独立な観測に基づいて考える.
- 2標本t統計量は次のよう,

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

2標本t検定に対するリサンプリングによる方法, パート 1

- $H_0: E(X) = E(Y)$と$H_a: E(X) \neq E(Y)$の検定を, Xからのn_X個の独立な観測とYからのn_Y個の独立な観測に基づいて考える.
- 2標本t統計量は次のよう,

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

- n_X, n_Yが大きいとき, TはH_0の下で近似的に$N(0,1)$に従う.

2標本t検定に対するリサンプリングによる方法, パート 1

- $H_0: E(X) = E(Y)$と$H_a: E(X) \neq E(Y)$の検定を, Xからのn_X個の独立な観測とYからのn_Y個の独立な観測に基づいて考える.
- 2標本t統計量は次のよう,

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

- n_X, n_Yが大きいとき, TはH_0の下で近似的に$N(0,1)$に従う.
- n_X, n_Yが小さいときにはTの理論的な帰無分布は分からない.

2標本t検定に対するリサンプリングによる方法, パート 1

- $H_0: E(X) = E(Y)$と$H_a: E(X) \neq E(Y)$の検定を, Xからのn_X個の独立な観測とYからのn_Y個の独立な観測に基づいて考える.
- 2標本t統計量は次のよう,

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

- n_X, n_Yが大きいとき, TはH_0の下で近似的に$N(0,1)$に従う.
- n_X, n_Yが小さいときにはTの理論的な帰無分布は分からない.
- 並べ替えやリサンプリングによる方法を用いてみよう.

2標本t検定に対するリサンプリングによる方法, パート 2

2標本t検定に対するリサンプリングによる方法，パート 2

1. 2標本t統計量Tをデータ$x_1, \ldots, x_{n_X}$と$y_1, \ldots, y_{n_Y}$に基づいて計算する.

2標本t検定に対するリサンプリングによる方法, パート 2

1. 2標本t統計量Tをデータ$x_1, \ldots, x_{n_X}$と$y_1, \ldots, y_{n_Y}$に基づいて計算する.
2. $b = 1, \ldots, B$(Bは大きな数, 例えば1,000)に対し, :

2標本t検定に対するリサンプリングによる方法, パート 2

1. 2標本t統計量Tをデータ$x_1, \ldots, x_{n_X}$と$y_1, \ldots, y_{n_Y}$に基づいて計算する.
2. $b = 1, \ldots, B$(Bは大きな数, 例えば1,000)に対し, :
    1. $n_X + n_Y$個の観測データをランダムに並べ替える.

2標本t検定に対するリサンプリングによる方法，パート 2

1. 2標本t統計量Tをデータ$x_1, \ldots, x_{n_X}$と$y_1, \ldots, y_{n_Y}$に基づいて計算する.
2. $b = 1, \ldots, B$(Bは大きな数, 例えば1,000)に対し, :
    1. $n_X + n_Y$個の観測データをランダムに並べ替える.
    2. その初めのn_X個を$x_1^*, \ldots, x_{n_X}^*$とし, 残りのn_X個を$y_1^*, \ldots, y_{n_Y}^*$とする.

2標本t検定に対するリサンプリングによる方法，パート 2

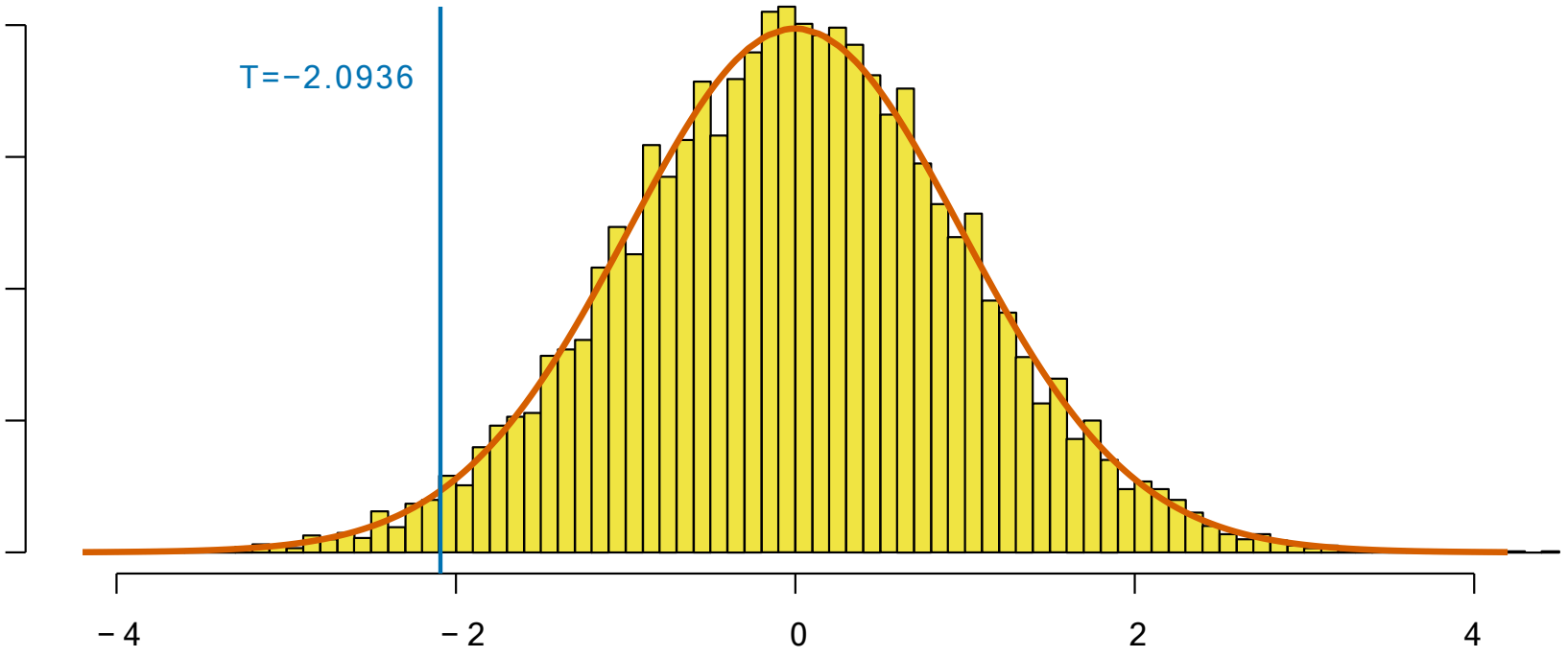
1. 2標本t統計量Tをデータ$x_1, \ldots, x_{n_X}$と$y_1, \ldots, y_{n_Y}$に基づいて計算する.
2. $b = 1, \ldots, B$(Bは大きな数, 例えば1,000)に対し, :
    1. $n_X + n_Y$個の観測データをランダムに並べ替える.
    2. その初めのn_X個を$x_1^*, \ldots, x_{n_X}^*$とし, 残りのn_X個を$y_1^*, \ldots, y_{n_Y}^*$とする.
    3. この並べ替えたデータに対して2標本t統計量を計算し, T^{*b}とする.

2標本t検定に対するリサンプリングによる方法，パート 2

1. 2標本t統計量Tをデータ$x_1, \ldots, x_{n_X}$と$y_1, \ldots, y_{n_Y}$に基づいて計算する.
2. $b = 1, \ldots, B$(Bは大きな数, 例えば1,000)に対し, :
    1. $n_X + n_Y$個の観測データをランダムに並べ替える.
    2. その初めのn_X個を$x_1^*, \ldots, x_{n_X}^*$とし, 残りのn_Y個を$y_1^*, \ldots, y_{n_Y}^*$とする.
    3. この並べ替えたデータに対して2標本t統計量を計算し, T^{*b}とする.
3. p値は次のように計算する.

$$\frac{\sum_{b=1}^{B} 1_{(|T^{*b}| \geq |T|)}}{B}.$$

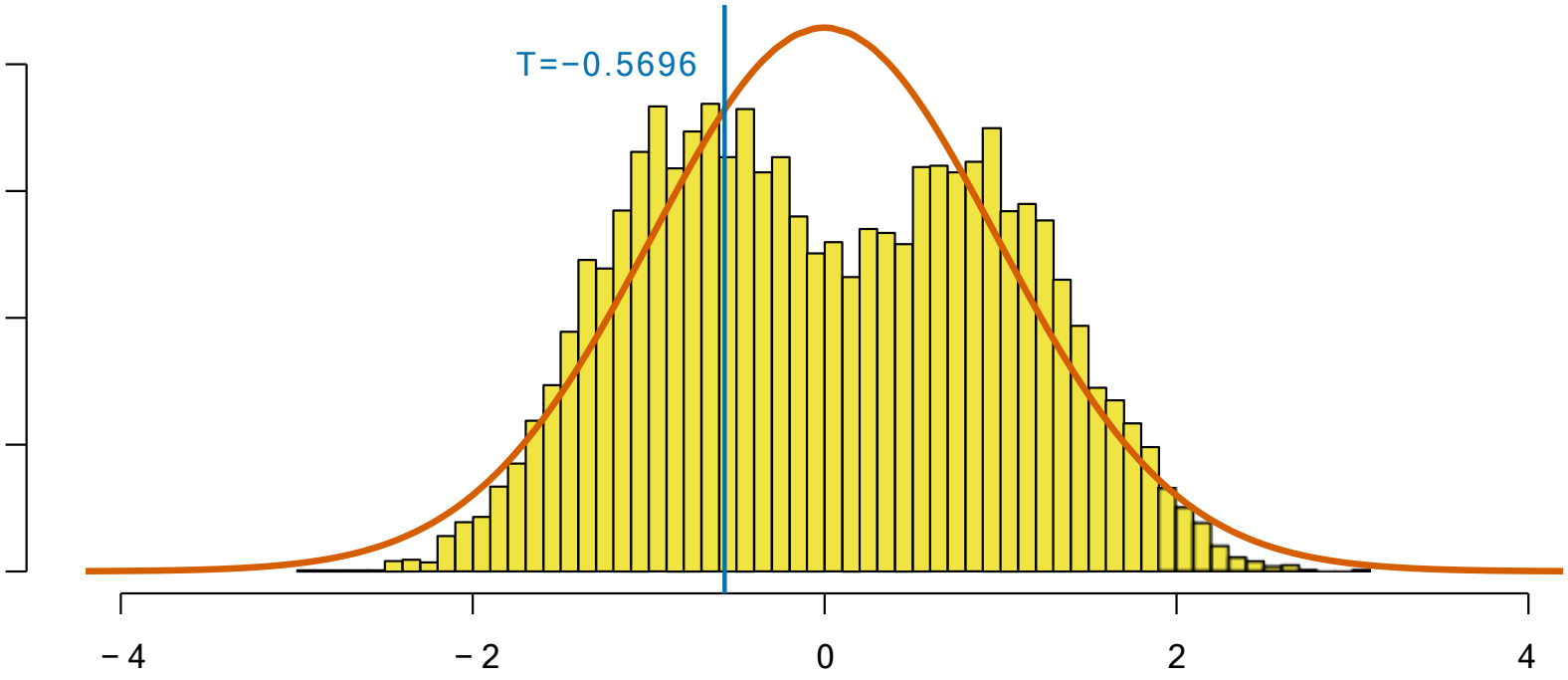
遺伝子データへの適用, パート 1



11番目の遺伝子に対する検定統計量の帰無分布

理論的なp値は0.041. リサンプリングによるp値は0.042.

遺伝子データに対する適用, パート 2



877番目の遺伝子に対する検定統計量の帰無分布

理論的なp値は0.571. リサンプリングによるp値は0.673.

リサンプリングによる方法について

リサンプリングによる方法について

- 理論的な帰無分布が利用できない場合や, 強い仮定が要求される場合にはリサンプリングによる方法が有用である. (常に有用となる!)

リサンプリングによる方法について

- 理論的な帰無分布が利用できない場合や, 強い仮定が要求される場合にはリサンプリングによる方法が有用である. (常に有用となる!)
- p値の計算におけるリサンプリングの方法を拡張してFDRのコントロールを行うこともできる.

リサンプリングによる方法について

- 理論的な帰無分布が利用できない場合や, 強い仮定が要求される場合にはリサンプリングによる方法が有用である. (常に有用となる!)
- p値の計算におけるリサンプリングの方法を拡張してFDRのコントロールを行うこともできる.
- 今回の例では2標本t検定について扱ったが, 他の検定統計量に対しても似たような方法を考えることができる.

日本語版への注

教科書ISLR2では統計モデルをしばしば利用しているが、統計モデルと統計的推測の説明が十分でないので,講義の理解に最低必要と思われる事項の概略を述べる。詳しくは数理統計学の教科書(例えば国友直人,「応用をめざす数理統計学」,朝倉書店;竹村彰通「現代数理統計学」,学術図書出版,久保川達也「現代数理統計学の基礎」,共立出版 など)を参照されたい。

1. 統計量・推定量
2. 積率法と最尤法
3. 条件付分布とベイズ法
4. 赤池情報量規準(AIC)
5. 最尤推定の漸近的性質
6. 区間推定と仮説検定

1. 統計量・推定量

実験・観察のデータをn個の実数を小文字を使って$(x_1, x_2, \ldots, x_n)$とする。データを母集団から標本抽出により得られるn個の独立標本を確率変数として理解するとき、大文字の記号を使って$(X_1, X_2, \ldots, X_n)$と表現する。標本を$(X_1, X_2, \ldots, X_n)$として、標本の関数を統計量(statistic)と呼ぶ。標本の関数である統計量により母集団を表現している未知母数を推測するとき、統計量を推定量(estimator)、推定量を使って標本として実際に観測値として得られるデータを代入した統計量の値を推定値(estimate)と区別する。
データがベクトルの場合には太文字$(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$,$(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$を利用することが多い。

母集団として確率分布の母数は未知であるパラメトリック・モデル(parametric model)では未知母数θ(ギリシャ文字のシータ)で表し、母数について母集団から(多くの場合には独立な)標本抽出により得られる標本$(X_1, X_2, \ldots, X_n)$から推測を行う。大きさnの標本の関数$\hat{\theta}(X_1, X_2, \ldots, X_n)$により母数を推定するとき関数$\hat{\theta}$(シータ・ハット)を推定量とする。$n$個の標本が観測値として$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$と云う値をとるときは推定値と呼ぶが、推定量$\hat{\theta}(X_1, X_2, \ldots, X_n)$の方は統計量で確率変数、推定値$\hat{\theta}(x_1, x_2, \ldots, x_n)$の方は数値, あるいはベクトルとなる。なお、ISLR2の議論では母数の存在を仮定しないノンパラメトリック(non-parametric)な統計モデルも多用されている。

2. 積率法と最尤法

大きさnの独立標本$X_1, X_2, \ldots, X_n$とするとき、標本から計算できる統計量は未知母数θの推定量として様々な方法が考えられる。例えばK. Pearsonは標本から計算される積率（標本積率）を母集団の積率の推定量に一致させる方法を提唱、例えば母集団の確率分布の期待値$\mu = E(X)$について標本平均 $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$, 母集団の確率分布の分散$\sigma^2 = E[(X - E(X))2]$には標本分散 s_n^2=$[1/(n-1)]\sum_{i=1}^{n}(X_i - \bar{X})^2$ を代入する方法であり、積率法（method of moments）と呼ばれている。推定の対象が母集団としての確率分布の積率であればこの方法を一般的に利用することが可能であるが、その他の場合には適用は困難となる。

母集団の確率分布を$p(x|\theta)$(密度関数なら$f(x|\theta)$)、$X_1, X_2, \ldots, X_n$をランダム・サンプリングによるn個の独立標本とする。このとき確率変数$X_1, X_2, \ldots, X_n$は互いに独立なので、離散確率分布の場合には同時確率関数

$$p(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

あるいは連続分布の場合には同時密度関数

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

と書ける。この関数を同時確率分布としてではなく母数θの関数とみて

$$L_n(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\theta)$$

尤度（likelihood）関数, $x_1, x_2, \ldots, x_n$を略して$L_n(\theta)$と表すこともある。母集団分布として連続型分布の場合も同様に独立標本の場合には尤度関数は

$$L_n(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

によって定められる。

尤度（likelihood）関数 $L_n(\theta|x_1, x_2, \ldots, x_n)$ を最大化するようなθを最尤推定値(maximum likelihood estimate)とよぶ。観測値 $x_1, x_2, \ldots, x_n$を標本(確率変数) $X_1, X_2, \ldots, X_n$で置き換えての尤度関数を最大にするような推定量$\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$を最尤推定量(maximum likelihood estimator)と呼ぶ。

n個の標本が互いに独立な場合には、対数尤度関数(log-likelihood)

$$l_n(\theta) = \log L_n(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

あるいは

$$l_n(\theta) = \log L_n(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

変数θで微分してゼロとおけば

$$\frac{\partial l_n(\theta)}{\partial \theta} = 0$$

となる。この方程式のように尤度関数や対数尤度関数を母数で微分してゼロと置いた方程式を尤度方程式(likelihood equation)と呼ばれている。

3. 条件付分布とベイズ法

母数θについて事前情報(prior information)が既知の事前分布 $p(\theta)$ として得られている場合には,
標本の条件付分布 $p(z|\theta)$（ここでzは観測可能な確率分布を持つ）と組み合わせると事後分布を
活用できる。事後分布にもとづき母数の推測を行う方法はベイズ法(Bayes method)と呼ばれている。

ベイズの定理(Bayes Theorem)とは連続分布の事前分布(prior distribution) $p(\theta)$, $\boldsymbol{z} = (z_1, \dots, z_n)$ の
条件付分布 $p(\boldsymbol{z}|\theta)$が与えられたもとで事後分布(posterior distribution)は

$$p(\theta|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\theta)p(\theta)}{p(\boldsymbol{z})}$$

で与えられる、という命題である。ただし $p(\boldsymbol{z}) = \int [p(\boldsymbol{z}|\theta)p(\theta)]d\theta$ とする。
確率変数が離散変数なら $p(\theta_i|\boldsymbol{z})\ (i = 1, \dots, m), p(\boldsymbol{z}) = \sum_{j=1}^{m} p\left(\boldsymbol{z}\middle|\theta_j\right)p(\theta_j)$とすれば同様の命題が
得られる。

母数θについての情報は事後分布に縮約されるので、例えば事前分布の事後平均(期待値),
事後中央値などが利用できる。一定の条件下で（例えば事前分布が適切であれば）、こうして得ら
れるベイズ解は理論的にも許容的(admissible, この解よりも未知母数について一様により優れた解
は存在しない)であることが知られている。

4. 赤池情報量規準(AIC)

Akaike (1973, 1974)

パラメター数kとすると, 尤度関数fから$-2\log f$のバイアスは近似的に$2k$となる. このバイアスを補正してAICは次のように定義される. このAICを最小化したモデルを選択する基準はAIC最小化規準と呼ばれている.

$$\mathrm{AIC} = -2\log f(x \mid \hat{\theta}) + 2k$$

k: 自由なパラメータ数

$\hat{\theta}$: 最尤推定量

$f(x|\hat{\theta})$: 最大対数尤度 $\quad f\left(x|\hat{\theta}\right) = \max_{\theta} f(x|\theta)$

慣例で「情報量基準」ではなく「情報量規準」と書かれる

5. 最尤推定の漸近的性質

標本数nがかなり大きい場合には「良い推定量」の議論を展開することは可能である。標本数が大きいときの統計理論は漸近理論、大標本理論と呼んでいる。標本数が大きければ最尤推定量は幾つかの正則条件の下で近似的（漸近的）に一致性、漸近正規性を持つことが知られている。

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \sim^a \ N(0, F_n),$$

ここでθ_0は真値で,$F_n^{-1} = \frac{1}{n} E\left[-\frac{\partial^2 l_n}{\partial^2 \theta} \,\middle|\, \theta = \theta_0 \right]$ はFisher情報量である。

このことは中心極限定理の拡張された内容と解釈できるが、実際的な観点からも最尤推定量は標本数がある程度多ければ近似的な意味でばらつきの少ない推定量なのでよい推定量となることは重要な意味がある。この議論は多次元にも拡張されている。

6. 区間推定と仮説検定

未知母数の値を一点として推定する方法が点推定(point estimation), ある信頼度の基準で未知母数が含まれる区間を標本から推定する方法は区間推定(interval estimation)と呼ばれる。
例：ある家電メーカーの生産するテレビの寿命が未知の平均μ ,標準偏差σ_0(例えば)=10で既知の正規分布$N(\mu\ ,{\sigma_0}^2)$にしたがっているとする。n個の独立標本$X_1, X_2, \ldots, X_n$が利用可能とするとき信頼係数99%の信頼区間を求める。
正規分布の母集団から独立に得られた標本平均 $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$とすると期待値 $E(\bar{\mathrm{X}}) = \mu, Var\ (\bar{\mathrm{X}}) = {\sigma_0}^2/n$ の正規分布となる。標本平均を基準化してかなめの量(pivot) $Z = \frac{\bar{\mathrm{X}} - \mu}{\sqrt{{\sigma_0}^2/\mathrm{n}}}$ とおけば、この確率変数の分布は標準正規分布$N(0,1)$にしたがう。正規分布の数表を使えば, $P(-2.58 < Z < 2.58) = .99$, あるいは $P(-1.96 < Z < 1.96) = .95$ となる。確率変数Zの定義を代入して整理すれば$P(\bar{\mathrm{X}} - 2.58\sqrt{{\sigma_0}^2/\mathrm{n}} < \mu < \bar{\mathrm{X}} + 2.58\sqrt{{\sigma_0}^2/\mathrm{n}}\ ) = .99$ となるので、母数μの99%信頼区間は $[\bar{\mathrm{X}} - 2.58\sqrt{{\sigma_0}^2/\mathrm{n}}, \bar{\mathrm{X}} + 2.58\sqrt{{\sigma_0}^2/\mathrm{n}}]$ で与えられる。このようにして得られた信頼区間は標本平均のみに依存しているので標本平均が母平均の情報を代表していると見なせ、区間の長さは $5.16\sqrt{{\sigma_0}^2/\mathrm{n}}$ であるので、標本数nが大きくなれば区間が狭くなりいわばより精度が高く区間推定できる。
母集団分布から大きさnの独立標本$X_1, X_2, \ldots, X_n$が得られる時、区間の上限を示す統計量$U(X_1, X_2, \ldots, X_n)$と下限を示す統計量$L(X_1, X_2, \ldots, X_n)$を構成し、2つの統計量を使って確率 $P(L(X_1, X_2, \ldots, X_n) < \theta < U(X_1, X_2, \ldots, X_n)) = 1 - \alpha$ によって信頼係数$100(1-\alpha)$を定めるとき区間$[L(X_1, X_2, \ldots, X_n), U(X_1, X_2, \ldots, X_n)]$を信頼区間(confidence interval)と呼ぶ。

既知の母集団の分布族の下で分布型を決める母数をθとして、範囲を大文字Θで表して母数空間と呼び、帰無仮説を母数空間の一部分Θ_0により$H_0: \theta \in \Theta_0$で表す。対立仮説は母数空間Θの別の一部分Θ_1により$H_1: \theta \in \Theta_1$で表す。

標本を$X = (X_1, X_2, \ldots, X_n)$とすると標本から仮説を検定する問題は標本がどのような領域に入れば仮説を棄却すべきかの問題に帰着できる。この棄却される領域を棄却域(rejection region)と呼び、R_nで表そう。この領域が条件

$$P(X \in Rn|\theta) = \alpha \quad (\text{任意の}\theta \in \Theta_0)$$

で決めれば、左辺の確率は帰無仮説が正しいときに標本が棄却域に入る確率を表す。このときには仮説が正しいにもかかわらず仮説を棄却することになるので誤った決定を下しているが、この誤りを犯す確率を第1種の過誤と呼ばれる。ここで$0 < \alpha < 1$をあらかじめ指定しておけばこの誤りを犯す確率をコントロールしていることになるが、伝統的にはこの値を1%,5%,10%などいう切りのよい値としている。この第一種の過誤100α%を固定して標本 Xから棄却域を決めてれば統計的な検定方式が定められる。

ここで統計的な検定方式を複雑にする問題は検定に際して誤りを犯す可能性としては既に説明した第一種の過誤ばかりではない。帰無仮説が正しくないときに仮説を受容する確率は

$$P(X \in R_n^c \,|\theta) \quad (\text{任意の}\theta \in \Theta_1)$$

と表し、第2種の過誤と呼ぶ。(ここで領域 R_n^c は領域 R_n の補集合を示す。)この第2種の過誤はやはり誤りなので出来るだけこの確率を小さくすることが望ましいが、この確率を小さくすることは確率

$$\beta_n(\theta) = P(X \in Rn|\theta)\ (\theta \in \Theta_1)$$

を大きくすることと同等である。この確率は対立仮説が正しいときに帰無仮説を棄却して対立仮説を受容する確率を意味するので検定の検出力(power)と呼んでいる。ここで左辺の$\beta_n(\theta)$ (ギリシャ文字のベータ)にはわざわざ未知母数θを変数としたが、一般には検出力は未知母数に依存するのでこの値を母数に無関係に小さくすることはできない。

日本語版のあとがき

このスライド講義録の原著者のトレバー・ヘイスティ(Trevor Hastie)教授とロバート・ティブシラニ(Robert Tibshirani) 教授（共にスタンフォード大学統計学科教授）は過去30年ほどの間に米国の統計学における大きな流れとなっている統計的学習理論の発展の重要な一翼を担ってきた研究者達である。統計的学習理論について数多くの業績があるが、特にHastie教授はGAM(一般化加法モデル）の開発、Tibshirani教授はLASSOの提唱、などで著名であり、最近でもDeep-learning(深層学習)の二重降下に関する重要な貢献などもある。さらにスライド(https://www.statlearning.com/)からも垣間見れるように、数理統計・計算統計の先端的な研究と共に生物・医学統計における実際の応用での統計的データ分析にも精通している。したがって、2023年時点においてISLR2（An Introduction to Statistical Learning with Applications in R, 第2版,2021年）は統計的学習理論に関して利用可能な最良な教科書と云っても過言ではないだろう。

統計数理研究所において2021年に開始した「統計エキスパート人材育成」計画では、このRによるISLR2を統計学教育の中心的な教材の一つとして利用している。このスライド日本語版が今後、日本における大学・大学院における統計学の専門教育の一助になれば幸いである。なお日本語版に誤りがあれば公開版を修正するので、コメントや指摘はkunitomo (アット・マーク.)ism.ac.jpまで投稿されたい。

2023年5月訳者：国友直人(統計数理研究所・特任教授), 趙宇(東京理科大学経営学部・助教) &
湯浅良太(統計数理研究所・助教)