

機械学習法

ATR脳情報解析研究所 計算脳イメージング研究室 室長
理研革新知能統合研究センター チームリーダー
CINET 客員研究員
大阪大学院 生命機能研究科 客員準教授

山下 宙人

おすすめ教材

Coursera machine learning by Prof. Ng

Pattern recognition and machine learning

4.9 ★★★★★
30343件のレビュー >

機械学習 からの人気レビュー

★★★★★ by VB • 10月 3日 2016
Everything is great about this course. Dr. Ng dumbs it down with the complex math involved. He explained everything clearly, slowly and softly. Now I can say I know something about Machine Learning

★★★★★ by PM • 7月 14日 2019
This course is amazing and covers most of the ML algorithms. I really liked that this course has emphasized math behind each technique which helps to choose the best algorithm while solving a problem.

講師



Andrew Ng

CEO/Founder Landing AI; Co-founder, Coursera; Adjunct Professor, Stanford University; formerly Chief Scientist, Baidu and founding lead of Google Brain



講義内容

1. 機械学習について
2. モデルの複雑さとオーバフィット
3. 情報漏洩
4. 機械学習のBMIへの応用: 脳波のパターン判別
5. まとめ

機械学習とは

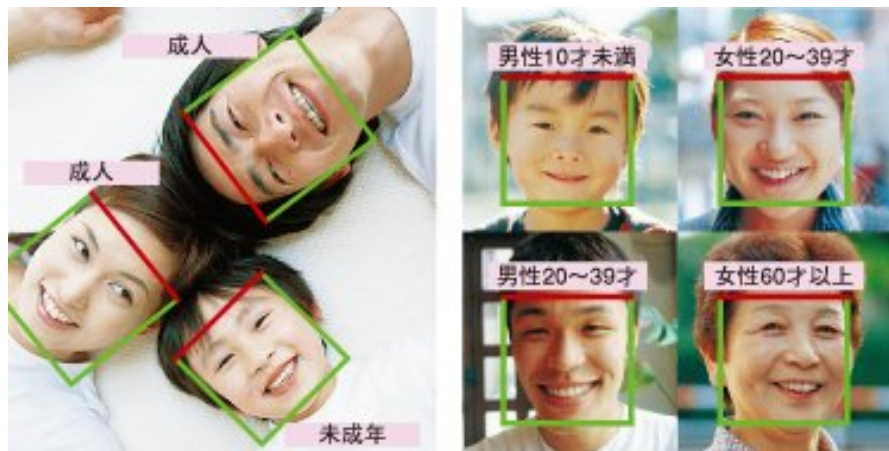
「コンピュータサイエンスの一研究分野で、明示的にプログラムしなくても学習する能力をコンピュータに与える」 (Arthur Samuel, 1959)

「コンピュータプログラムが、ある種のタスクTと評価尺度Pにおいて、経験Eから学習するとは、タスクTにおけるその性能をPによって評価した際に、経験Eによってそれが改善されている場合である」
(Tom Michel)

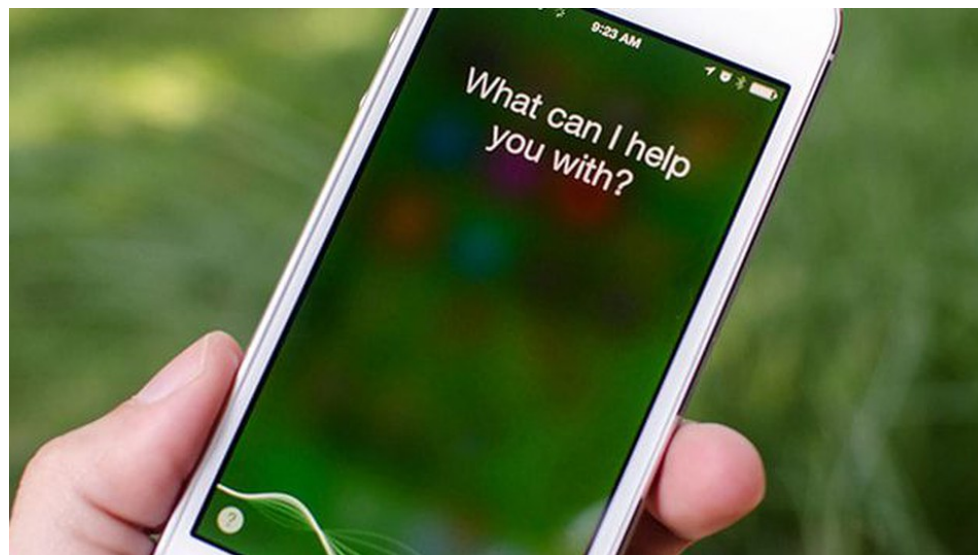
「機械学習とは、データから反復的に学習し、そこに潜むパターンを見つけ出すことです。そして学習した結果を新たなデータにあてはめることで、パターンにしたがって将来を予測することができます。」
(SAS)

機械学習のアプリケーション

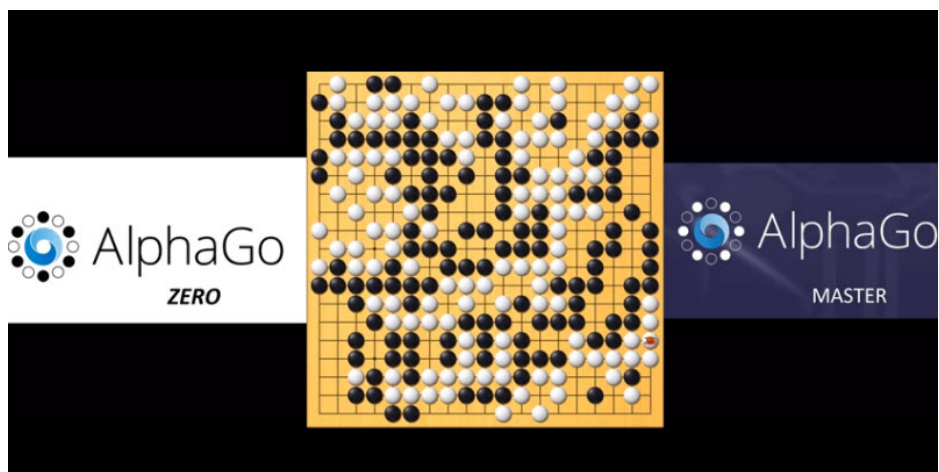
顔検知, 画像認識



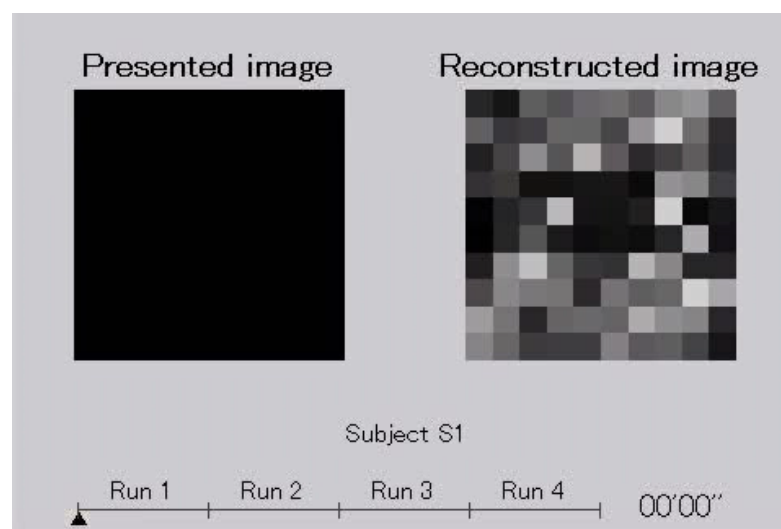
音声認識, 音声翻訳



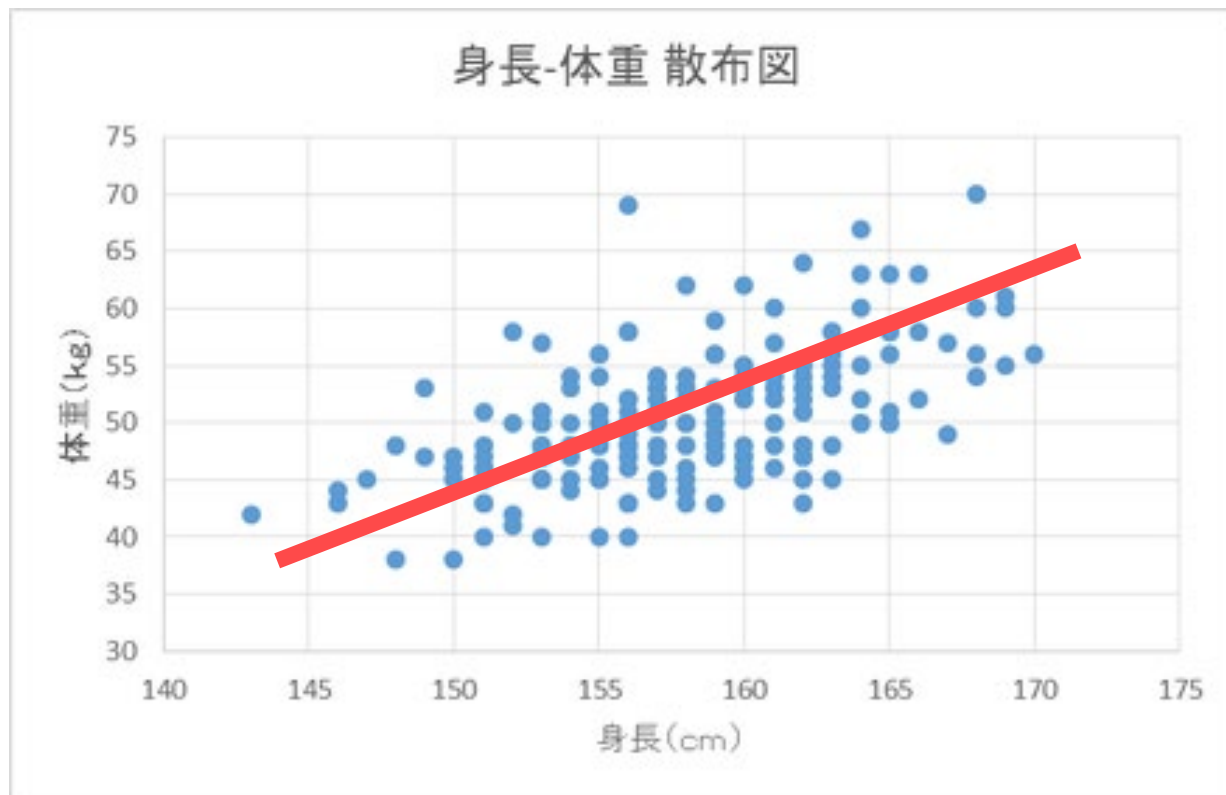
ボードゲーム



脳情報解読



機械学習のアプリケーション

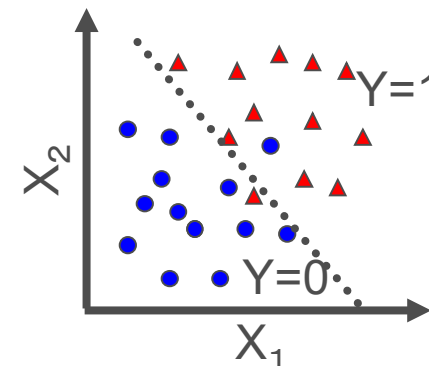


$$y = a_0 + a_1x$$

- **教師あり学習 (supervised learning)**

『入力 X と出力 Y の組から、入力 X から出力 Y を予測するモデルを学習する。』

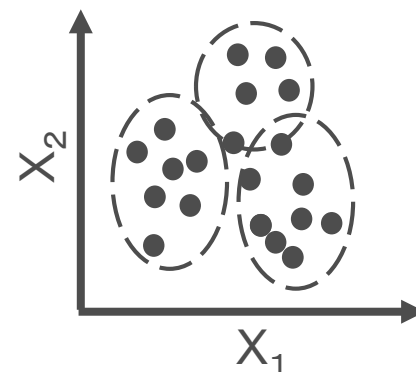
ex. 回帰、判別



- **教師無し学習 (un-supervised learning)**

『入力 X の隠れた構造を学習する。』

ex. クラスタリング、次元縮約 (主成分分析、独立成分分析)



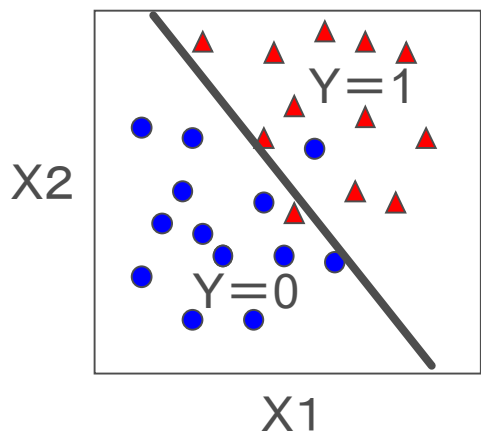
アドバンスドな問題: 強化学習、半教師あり学習、転移学習、マルチタスク学習 などなど

教師あり機械学習の問題分類

$$Y = f(X)$$

教師信号 特徴量

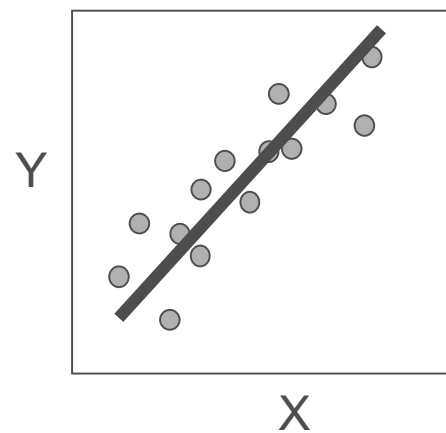
判別問題
(classification)



教師信号がカテゴリカルな値

ex. 物体認識

回帰問題
(regression)



教師信号が連続値

ex. 株価予測

教師あり機械学習の目的は関数近似と予測

例示データから関数 $f: X \rightarrow Y$ を学習する。

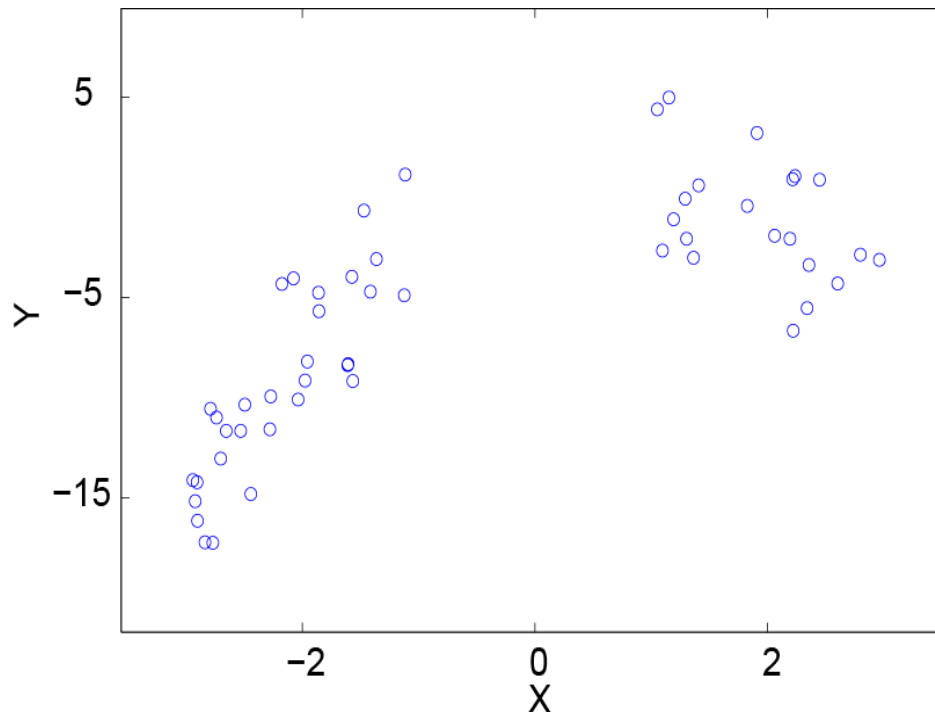
例示データに含まれない未知のデータに対しても予測できるように学習したい。

教師あり機械学習の目的は関数近似と予測

例示データから関数 $f: X \rightarrow Y$ を学習する。

例示データに含まれない未知のデータに対しても予測できるように学習したい。

問題 : 20次までの多項式を使って、 Y をよく説明する
 X の式を求めよ。



(パラメトリック)モデル

$$f(x; \theta) = a_0 + a_1x + \dots + a_{20}x^{20}$$

パラメータ

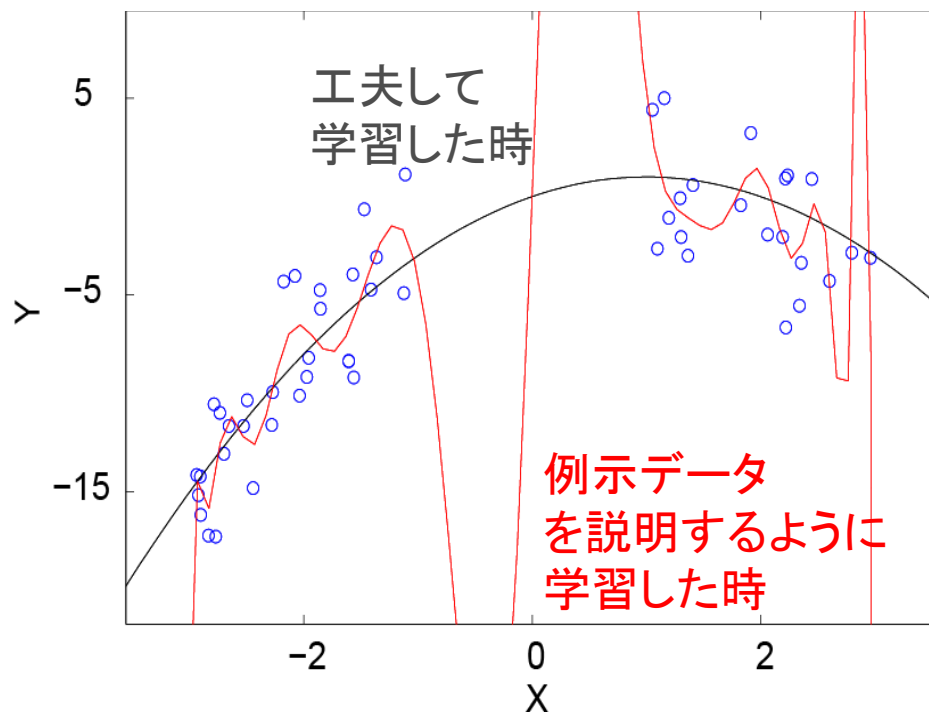
$$\theta = (a_0, \dots, a_{20})$$

教師あり機械学習の目的は関数近似と予測

例示データから関数 $f: X \rightarrow Y$ を学習する。

例示データに含まれない未知のデータに対しても予測できるように学習したい。

問題 : 20次までの多項式を使って、 Y をよく説明する
 X の式を求めよ。



(パラメトリック)モデル

$$f(x; \theta) = a_0 + a_1x + \dots + a_{20}x^{20}$$

パラメータ

$$\theta = (a_0, \dots, a_{20})$$

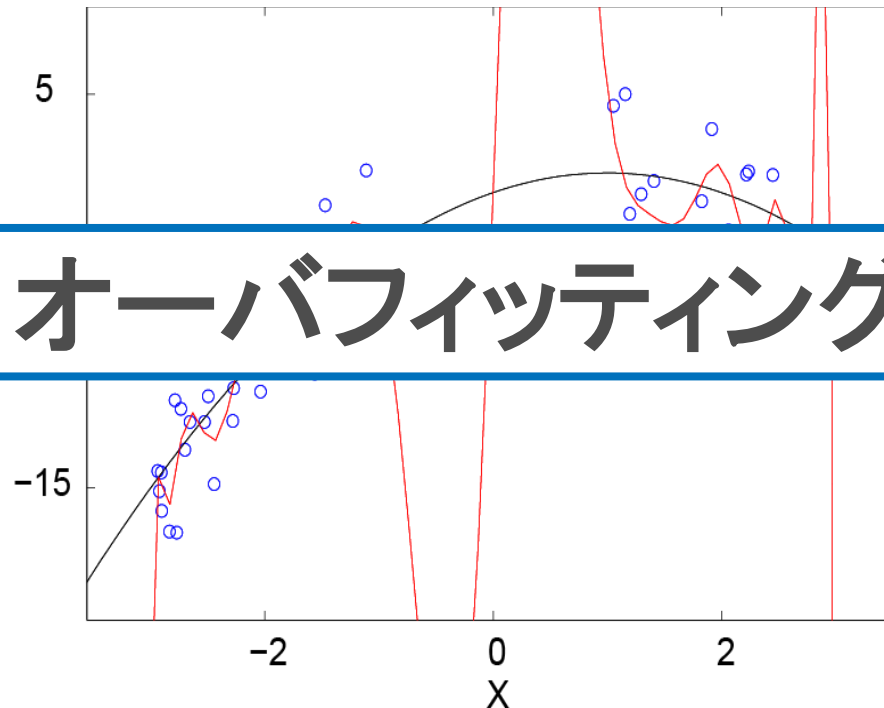
教師あり機械学習の目的は関数近似と予測

例示データから関数 $f: X \rightarrow Y$ を学習する。

例示データに含まれない未知のデータに対しても予測できるように学習したい。

問題 : 20次までの多項式を使って、 Y をよく説明する X の式を求めよ。

オーバフィッティング



(パラメトリック)モデル

$$f(x; \theta) = a_0 + a_1x + \dots + a_{20}x^{20}$$

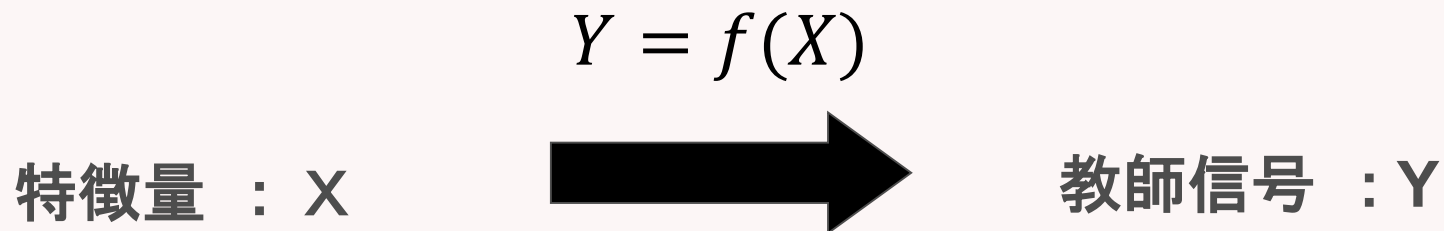
パラメータ

$$\theta = (a_0, \dots, a_{20})$$

- モデルの複雑さとオーバフィット・アンダーフィット。
- 情報漏洩の問題。

講義内容

1. 機械学習について
- 2. モデルの複雑さとオーバフィット**
3. 情報漏洩
4. 機械学習のBMIへの応用: 脳波のパターン判別
5. まとめ



特徴量と教師信号のペアからなるサンプル $(X, Y) = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ を用いて、関数 f を学習する問題。

$$Y = f(X; \Theta)$$

特徴量 : X



教師信号 : Y

特徴量と教師信号のペアからなるサンプル $(X, Y) = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ を用いて、関数 f を学習する問題。

関数を $f(X; \Theta)$ のように**パラメータ** Θ を用いて記述したものを**パラメトリックモデル**と呼ぶ。

$$Y = f(X; \Theta)$$

特徴量 : X



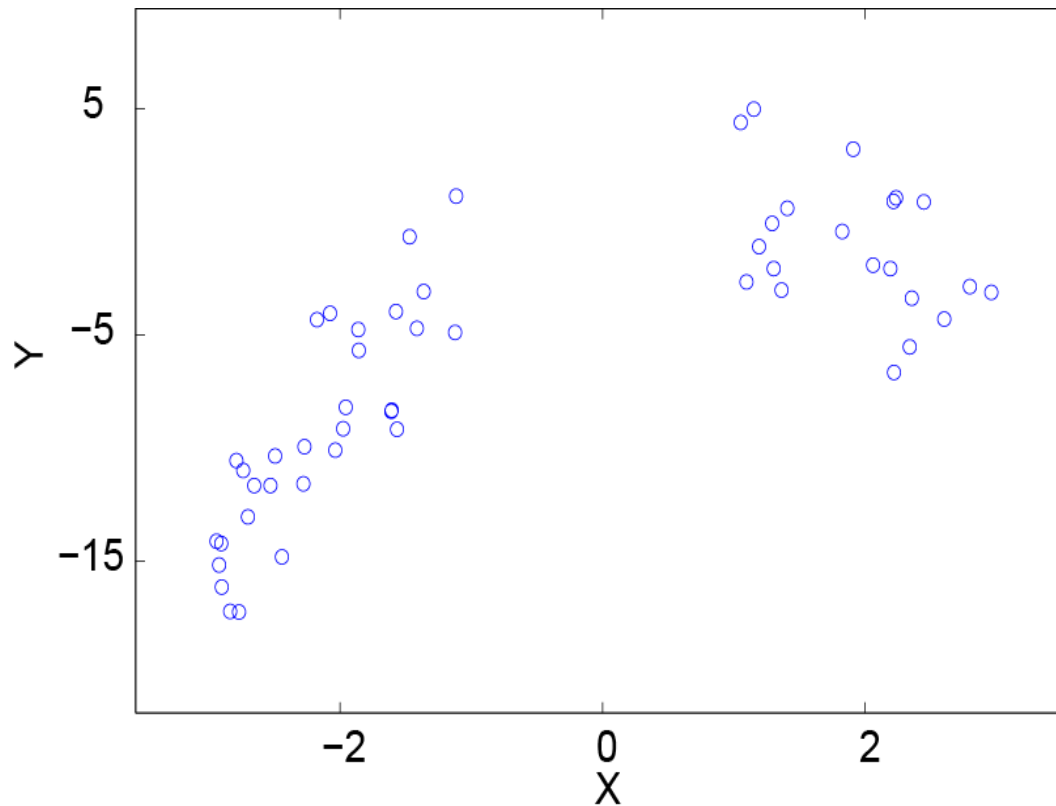
教師信号 : Y

特徴量と教師信号のペアからなるサンプル $(X, Y) = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$ を用いて、関数 f を学習する問題。

関数を $f(X; \Theta)$ のように**パラメータ** Θ を用いて記述したものを**パラメトリックモデル**と呼ぶ。

カーブフィッティング問題

問題 : 20次までの多項式を使って、Yをよく説明するXの式を求めよ。



(パラメトリック)モデル

$$f(x; \theta) = a_0 + \dots + a_{20}x^{20}$$

パラメータ

$$\theta = (a_0, \dots, a_{20})$$

モデルの複雑さ

次数の高いモデルほど、複雑な曲線が表現できる
～ パラメータの多いモデル = 複雑なモデル

次数1の多項式

$$y = a_0 + a_1x$$

次数2の多項式

$$y = a_0 + a_1x + a_2x^2$$

⋮
⋮
⋮

次数20の多項式

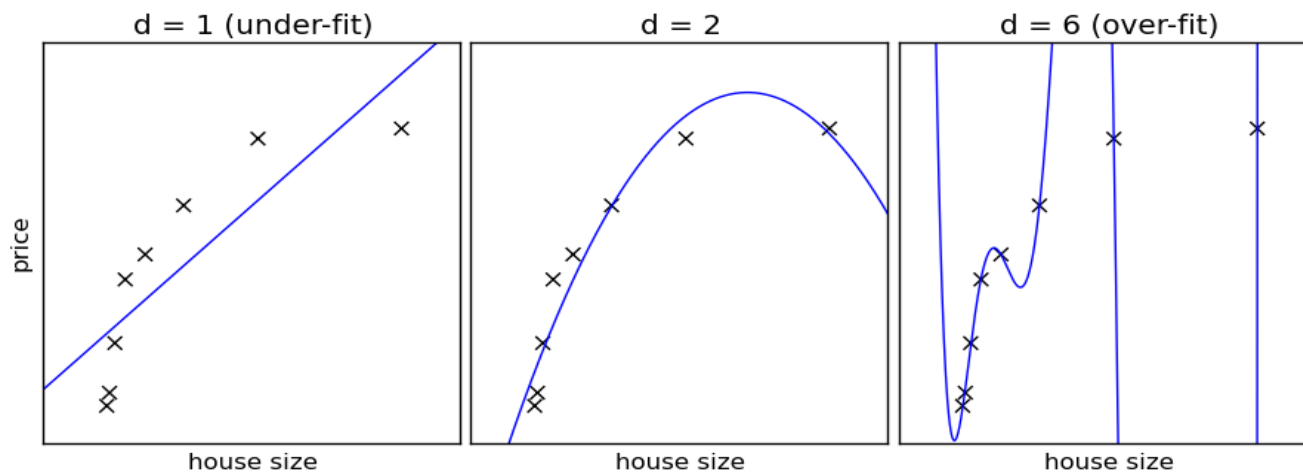
$$y = a_0 + a_1x + a_2x^2 + \cdots + a_{20}x^{20}$$

単純なモデル

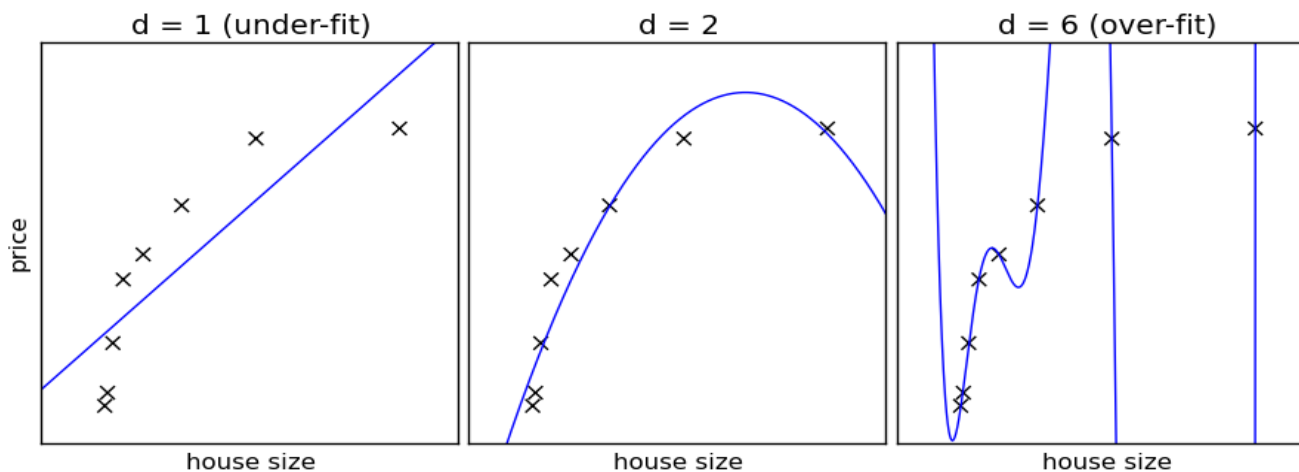
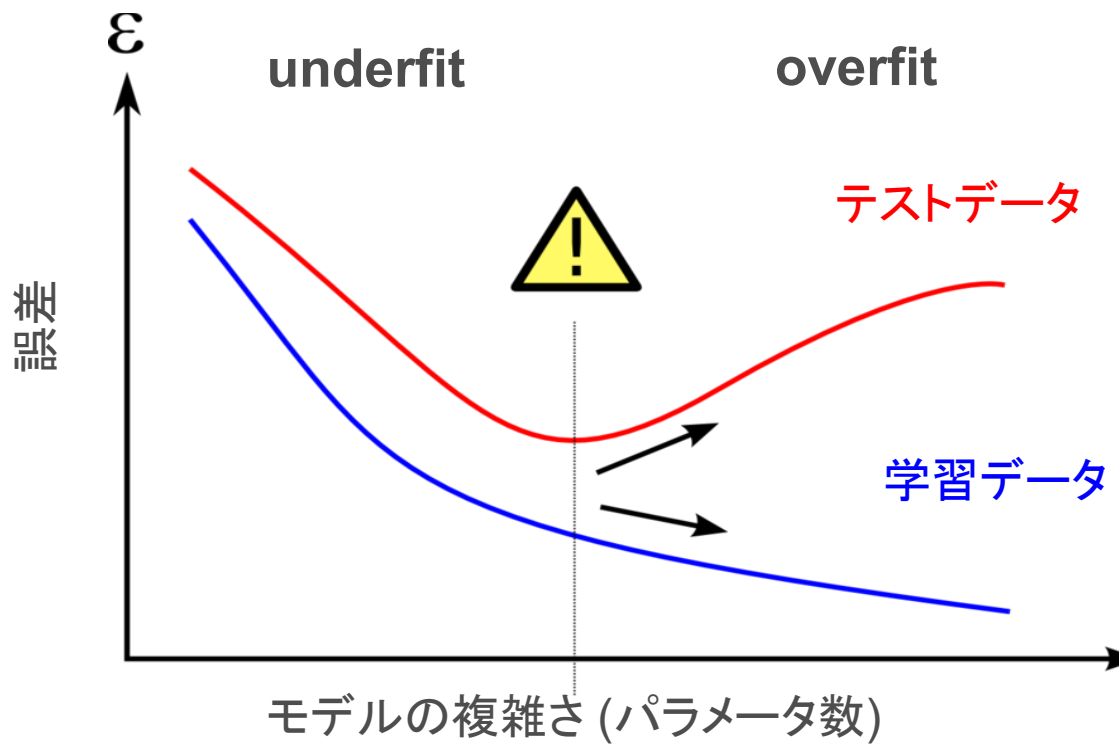


複雑なモデル

複雑なモデルが最適か？



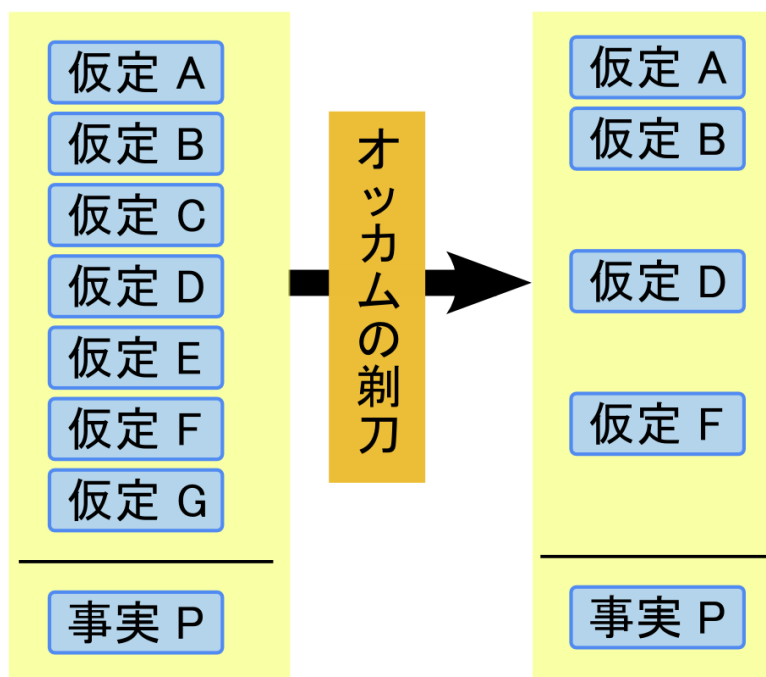
複雑なモデルが最適とは限らない



どうやったら適切な複雑さのモデルを選べるか？

モデルの複雑さの学習原理: オッカムのカミソリの原理 (けちの原理)

必要が無いなら多くのものを定立してはならない。少数の論理でよい場合は多数の論理を定立してはならない。— オッカム



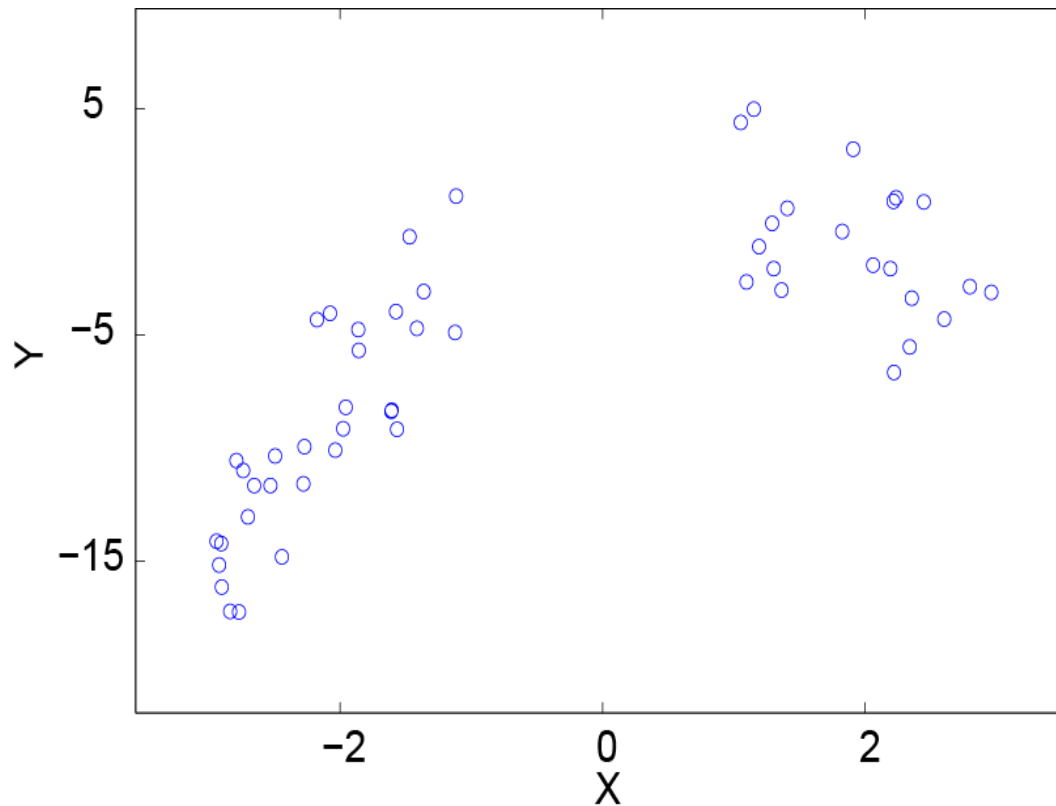
学習サンプルを同程度説明する2つのモデルがあるならばシンプルの方が良い。

モデルの複雑さの学習・3つのアプローチ

1. モデル選択
モデル選択基準と呼ばれる統計的な基準をもとに選ぶ
2. 正則化
パラメータを学習するときにパラメータに対して制約条件を課す
3. 特徴選択
問題分野の先行研究に基づいてモデルの複雑さをあらかじめ決める

カーブフィッティング問題

問題 : 20次までの多項式を使って、Yをよく説明するXの式を求めよ。



(パラメトリック)モデル

$$f(x; \theta) = a_0 + \dots + a_{20}x^{20}$$

パラメータ

$$\theta = (a_0, \dots, a_{20})$$

1. モデル選択による解法



Akaike, 1973

予測性能を測る統計的指標

$$\text{AIC} = (\text{二乗誤差}) + (\text{パラメータ数})$$

学習誤差項にモデルの複雑さに応じたペナルティを課す

| | | 学習二乗誤差 | ペナルティ | AIC | 単純なモデル |
|----------|---|--------|-------|------|-------------|
| 次数1の多項式 | ➡ | 20 | +1 | 21 | ↓ 複雑なモデル |
| 次数2の多項式 | ➡ | 15 | +2 | 17 | |
| 次数3の多項式 | ➡ | 14.5 | +3 | 17.5 | |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| 次数20の多項式 | ➡ | 10 | +20 | 30 | |

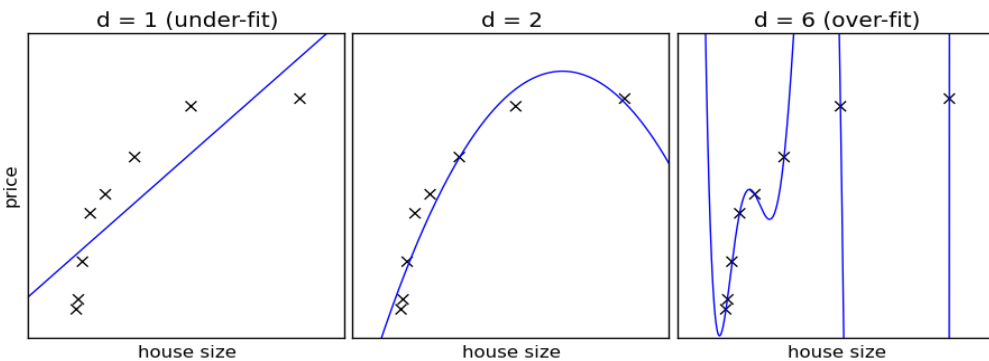
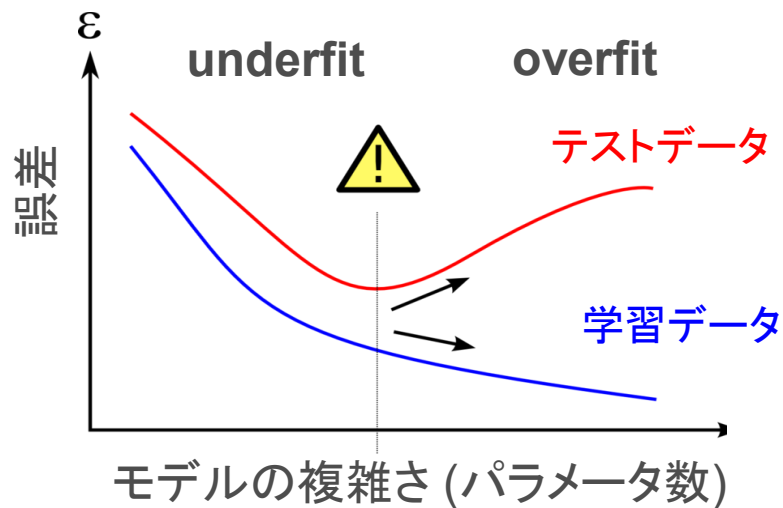
1. モデル選択による解法補足：AIC基準の正確な定義

$$\text{AIC} = -2 \times (\text{対数尤度}) + 2 (\text{パラメータ数})$$

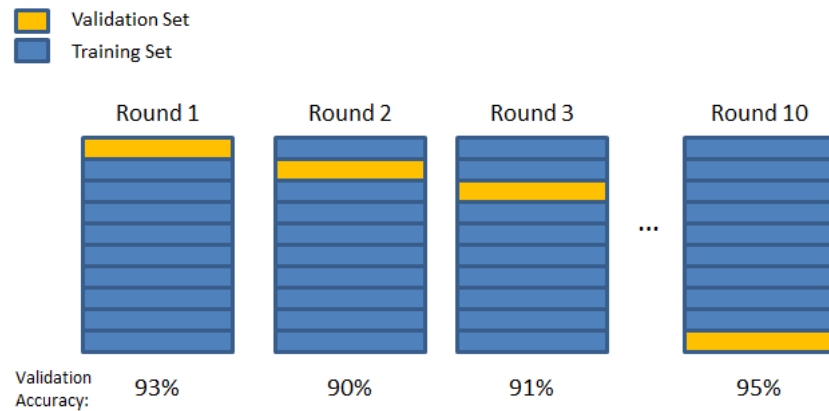
Akaike, 1973

- AICは正則なモデルにおいて期待対数尤度を漸近値を理論的に計算することによる導出される
- 入れ子構造のモデル群
- その後、様々なモデル選択基準 (ABIC, BIC, DIC, ...) が提案されている。

1. モデル選択による方法：クロスバリデーション誤差最小化



クロスバリデーション法



Final Accuracy = Average(Round 1, Round 2, ...)

2. 正則化による解法

$$\mathbf{a}^t \mathbf{x} = a_0 + \dots + a_{20} x^{20}$$

最小二乗法

$$E(\mathbf{a}) = \underbrace{\|y - \mathbf{a}^t \mathbf{x}\|^2}_{\text{フィッティング}}$$

L2-norm正則化

$$E(\mathbf{a}) = \underbrace{\|y - \mathbf{a}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \underbrace{\lambda \|\mathbf{a}\|^2}_{\text{制約}} \quad \text{リッジ回帰}$$

L1-norm正則化

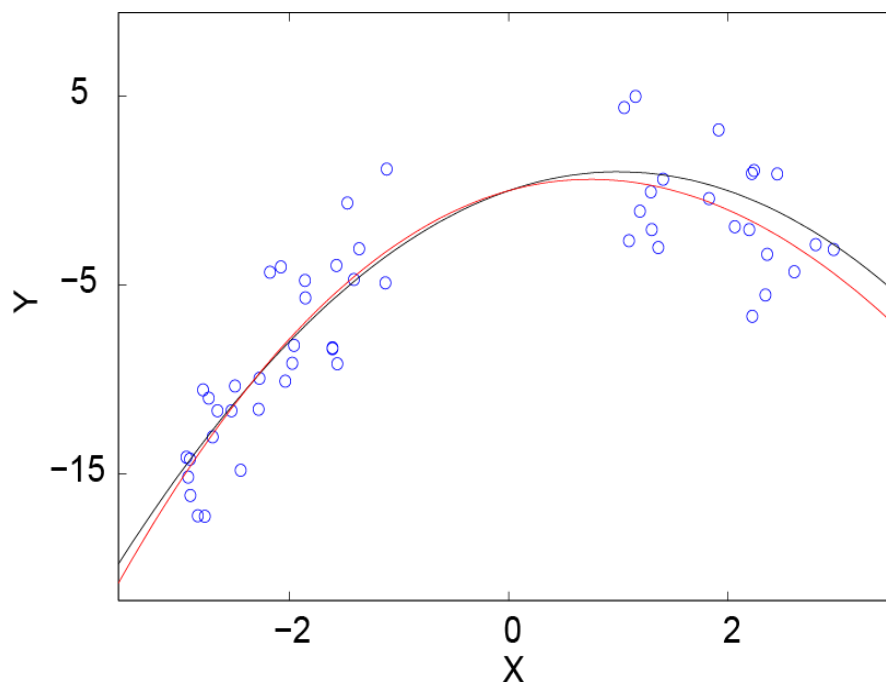
$$E(\mathbf{a}) = \underbrace{\|y - \mathbf{a}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\sum_i |a_i|}_{\text{制約}} \quad \text{LASSO}$$

λ : 正則化パラメータ ~ フィッティングと制約のバランスを決めるパラメータ

2. 正則化による解法

L1-norm正則化のもと学習した結果

$$y_n = 0.02 - 1.11x_n + 1.63x_n^2 + 0 \cdot x_n^3 \cdots + 0 \cdot x_n^{20}$$



スパース推定

= 自動的に重要なパラメータを選択しその値を推定

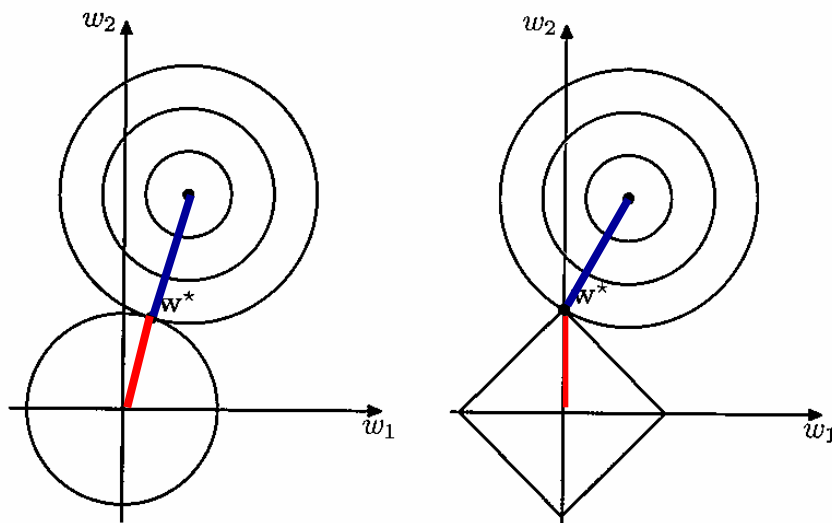
2. 正則化法：L1 norm 正則化とスパース推定

L2-norm正則化

$$E(\mathbf{w}) = \underbrace{\|y - \mathbf{w}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\|\mathbf{w}\|^2}_{\text{制約}}$$

L1-norm正則化

$$E(\mathbf{w}) = \underbrace{\|y - \mathbf{w}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\sum |w_i|}_{\text{制約}}$$



L2-norm

L1-norm

講義内容

1. 機械学習について
2. モデルの複雑さとオーバフィット
- 3. 情報漏洩**
4. 機械学習のBMIへの応用:脳波のパターン判別
5. まとめ

When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging

Robert Whelan and Hugh Garavan

予測精度のインフレが起こる時

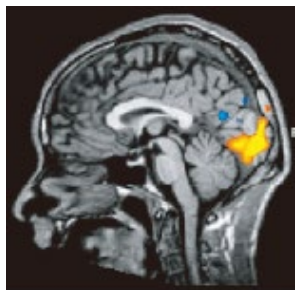
モデルの学習時に評価用データが含まれるとき

1. 判別関数の重み計算時には、学習データと評価用データをわけましょう。
2. アルゴリズムのパラメータ選択が必要な時は、選択のためのデータと評価用のデータはわけましょう。
3. 情報漏えいにはくれぐれも気をつけましょう。

1. 判別関数の重み計算時には、
学習データと評価用データをわけましょう。

ニューロイメージングデータからの 病態診断

0. 実験・計測



- イメージングデータ : I



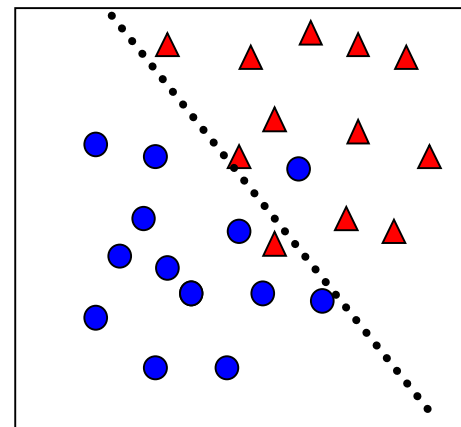
- 病態ラベル : y

1. 特徴量計算



- 画像前処理
- 変数の選択

2. 判別分析

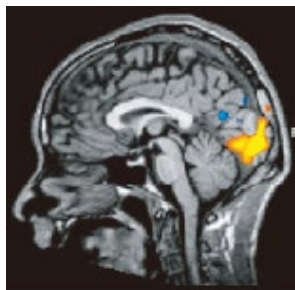


特徴量から病態を
予測するモデル

$$y = f(x)$$

ニューロイメージングデータからの 病態診断

0. 実験・計測



- イメージングデータ : I



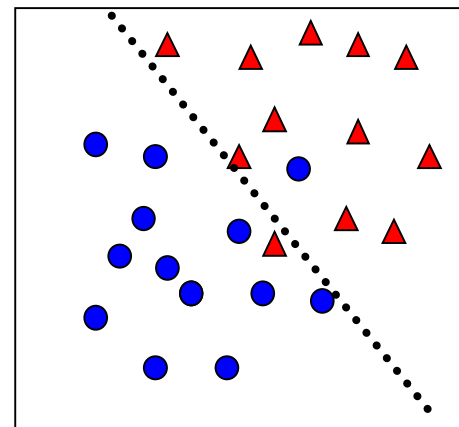
- 病態ラベル : y

1. 特徴量計算



- 画像前処理
- 変数の選択

2. 判別分析



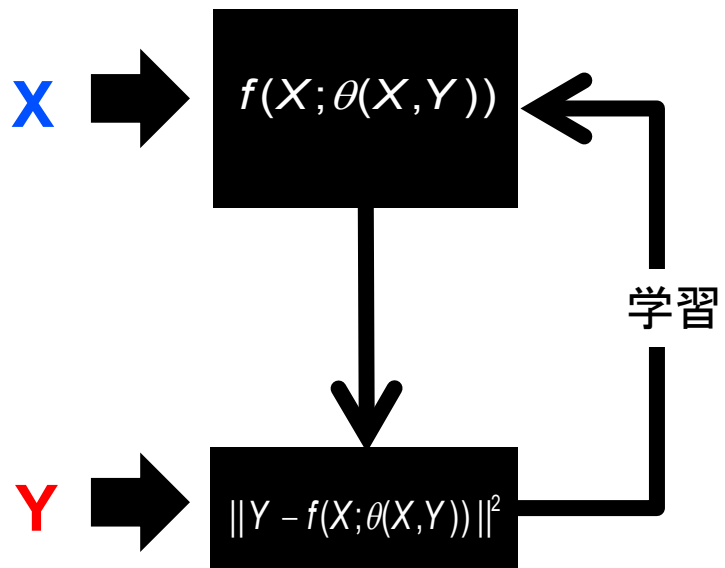
特徴量から病態を
予測するモデル

$$y = f(x)$$

予測性能を学習データで評価すると インフレが起こる

モデルの学習

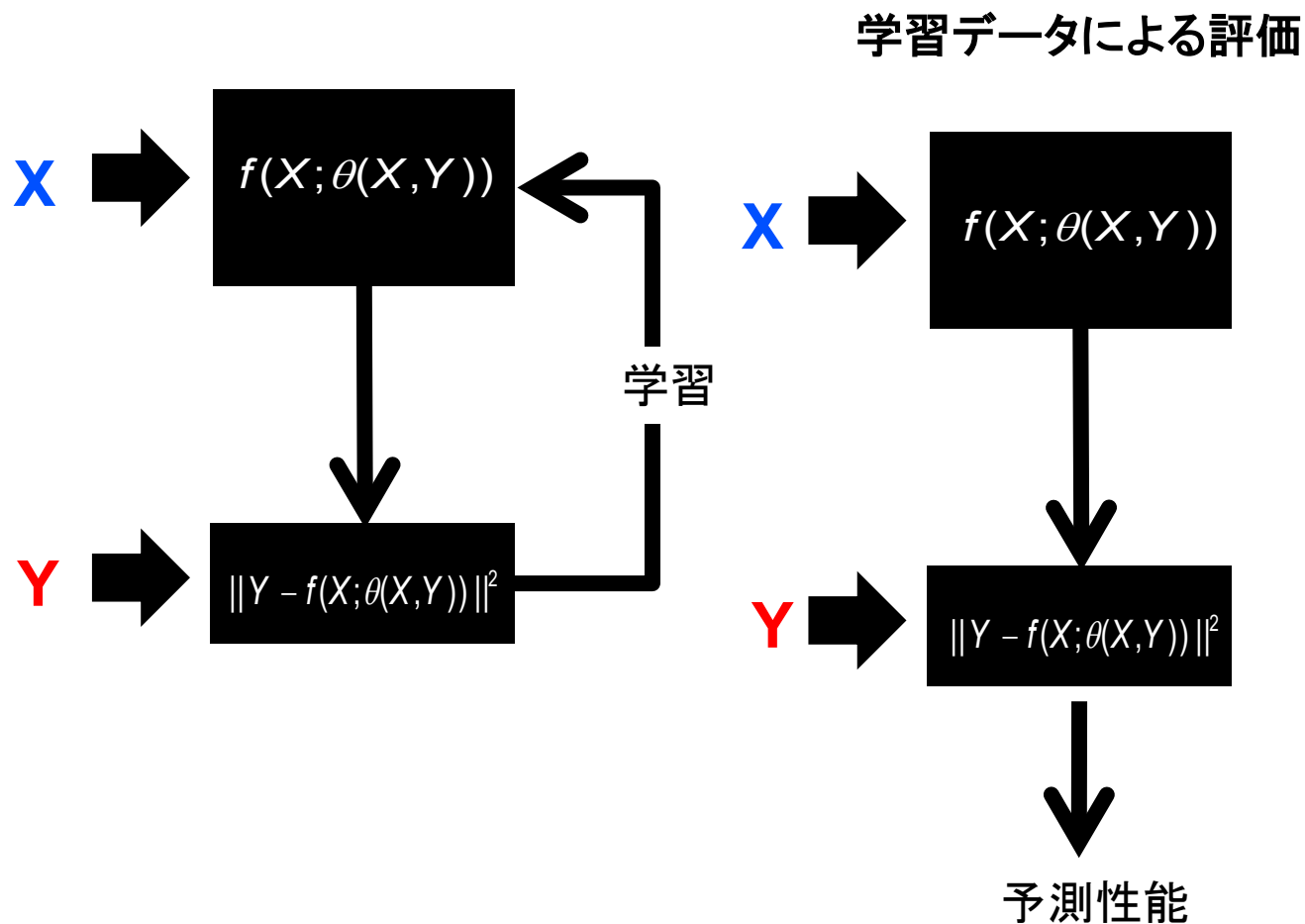
モデルの評価



予測性能を学習データで評価すると インフレが起こる

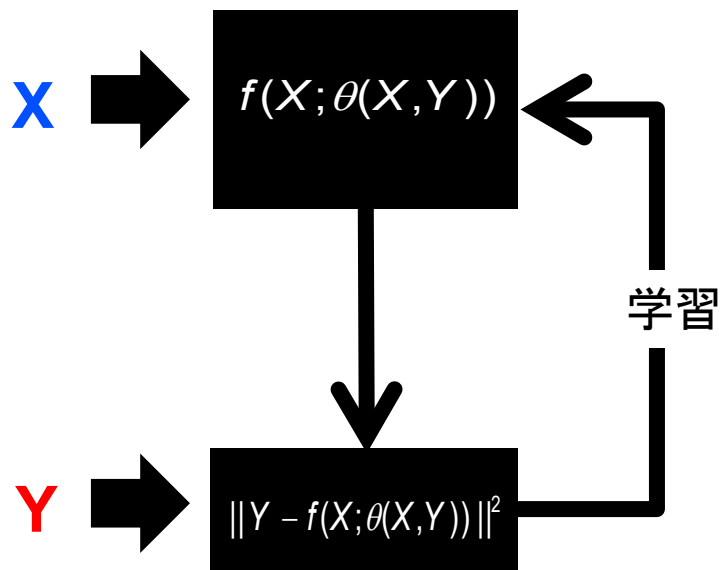
モデルの学習

モデルの評価



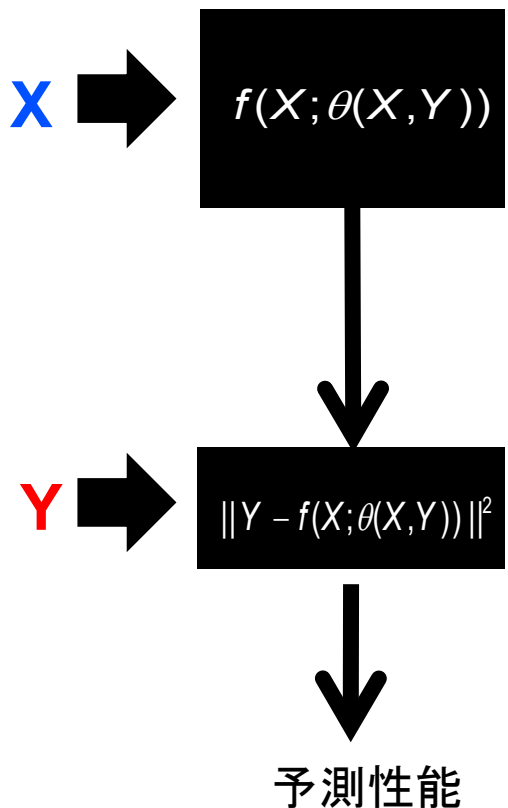
予測性能を学習データで評価すると インフレが起こる

モデルの学習

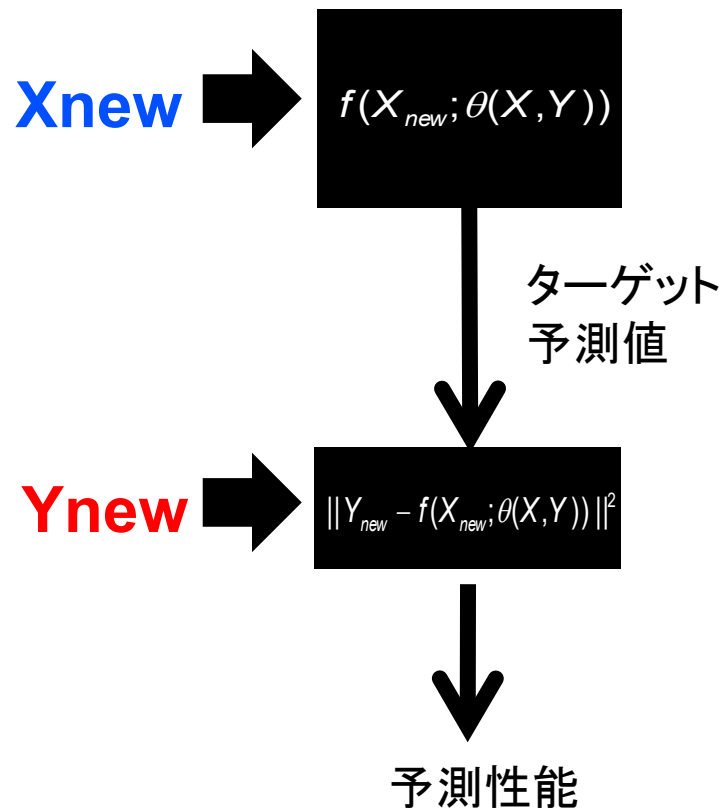


モデルの評価

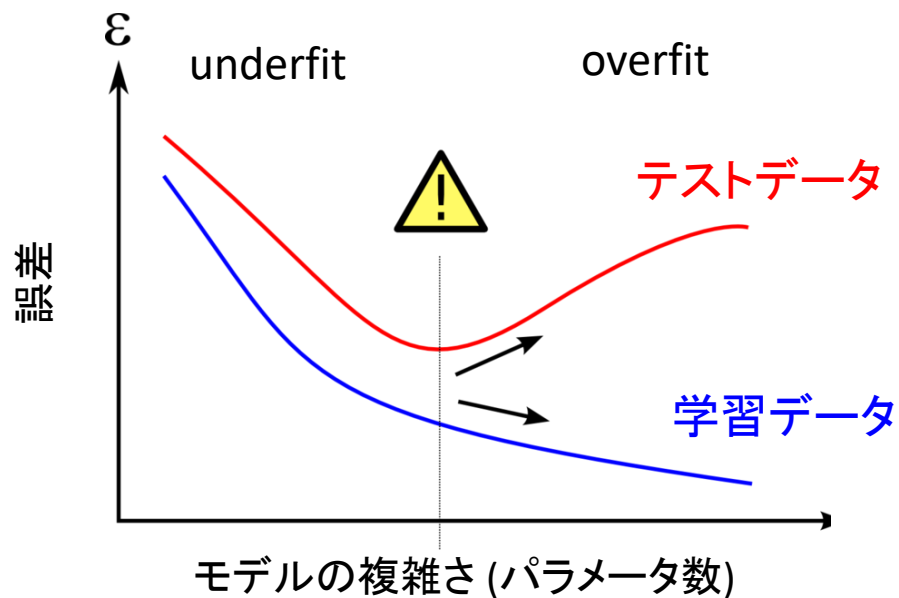
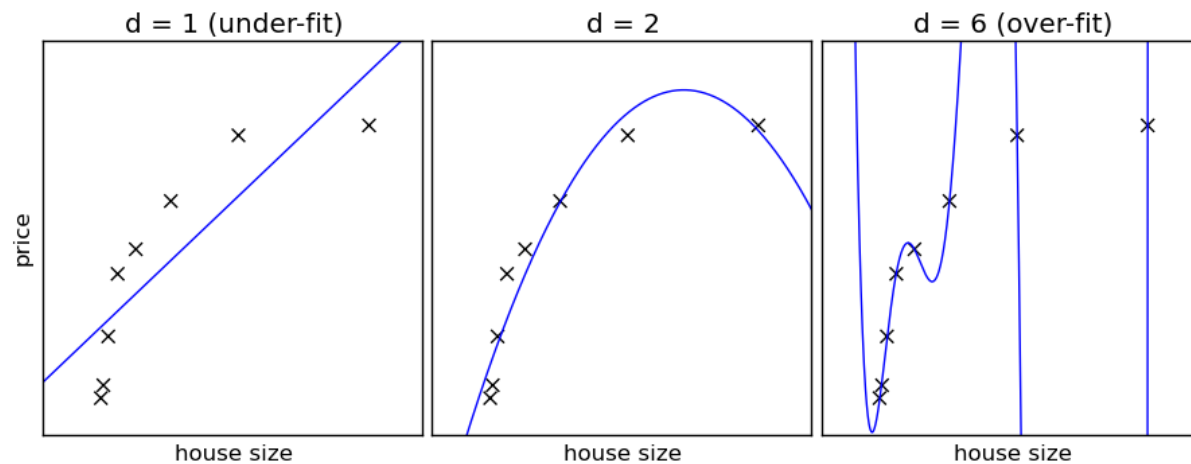
学習データによる評価



独立データによる評価

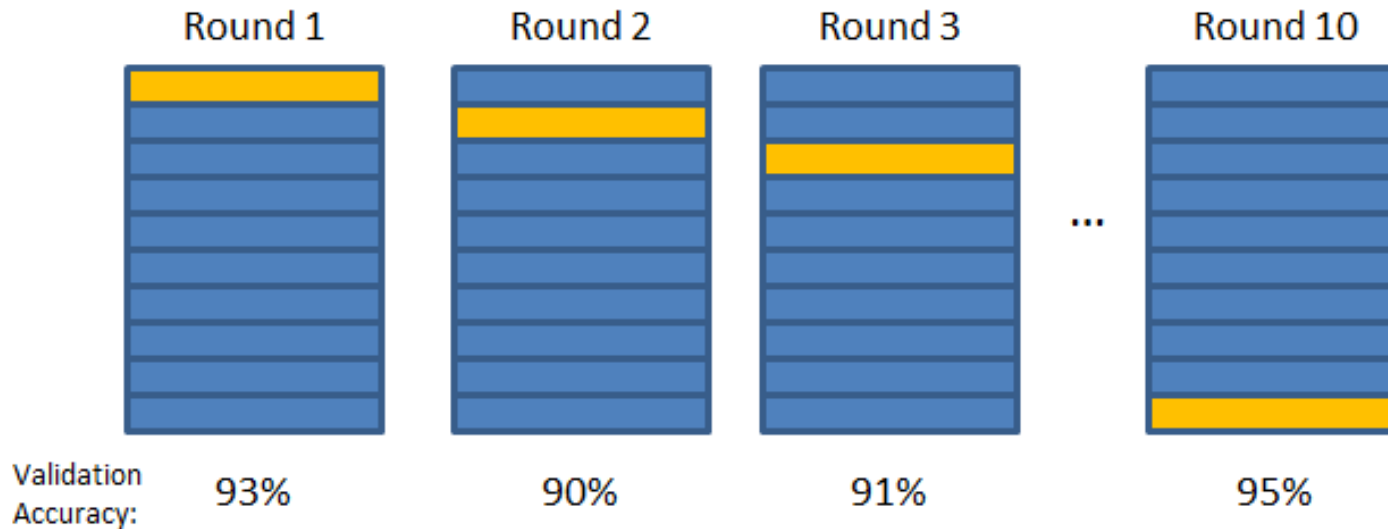
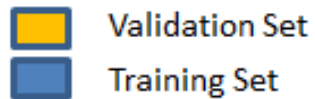


予測性能を学習データで評価すると インフレが起こる



予測性能を評価する方法

Cross validation



Final Accuracy = Average(Round 1, Round 2, ...)

その他の方法

- holdout method
- repeated holdout method
- bootstrap method

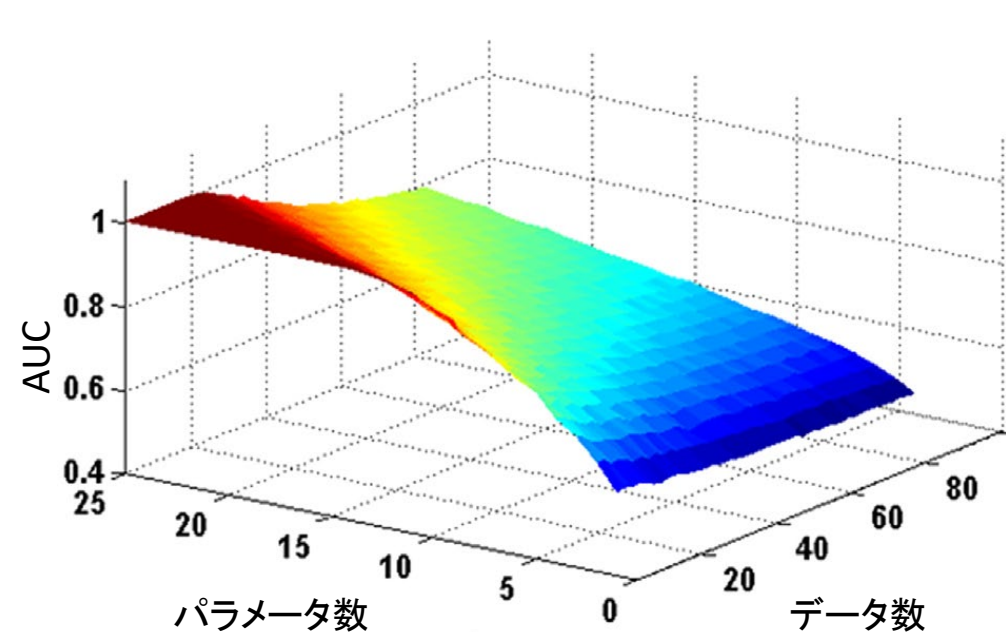
Overfitting

ランダム実験による例

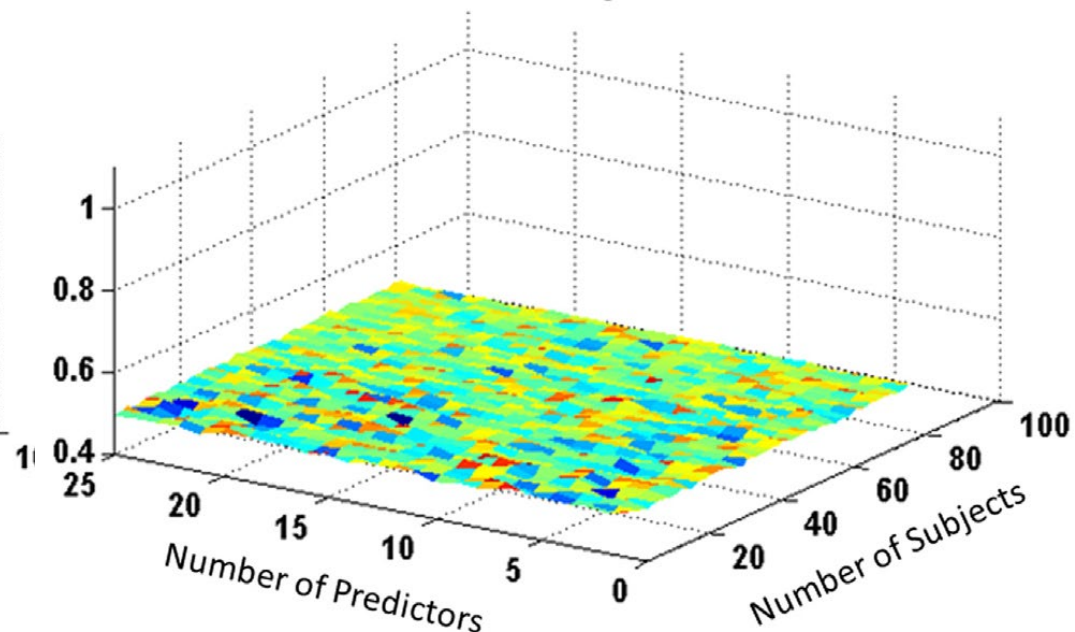
ターゲット変数と予測変数をランダムに設定。予測変数の数、データサンプル数を変える。

結果

- 理論上 0.5 のはずが、学習データでは非常に高い予測性能。
- パラメータ数が大きい時、サンプル数が小さいときに顕著。



学習データで評価



クロスバリデーションで評価

2. アルゴリズムのパラメータ選択が必要な時は、選択のためのデータと評価用のデータはわけましょう。

2. 正則化による解法

$$\mathbf{a}^t \mathbf{x} = a_0 + \dots + a_{20} x^{20}$$

最小二乗法

$$E(\mathbf{a}) = \underbrace{\|y - \mathbf{a}^t \mathbf{x}\|^2}_{\text{フィッティング}}$$

L2-norm正則化

$$E(\mathbf{a}) = \underbrace{\|y - \mathbf{a}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \underbrace{\lambda \|\mathbf{a}\|^2}_{\text{制約}} \quad \text{リッジ回帰}$$

L1-norm正則化

$$E(\mathbf{a}) = \underbrace{\|y - \mathbf{a}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\sum_i |a_i|}_{\text{制約}} \quad \text{LASSO}$$

λ : 正則化パラメータ ~ フィッティングと制約のバランスを決めるパラメータ

アルゴリズムパラメータの選択が 必要なケース

重みの学習

途中評価

最終評価

LASSO $\lambda=1 \rightarrow \theta_1(\mathbf{x}, y) \rightarrow p_1(\mathbf{x}_{te}, y_{te})$
LASSO $\lambda=2 \rightarrow \theta_2(\mathbf{x}, y) \rightarrow p_2(\mathbf{x}_{te}, y_{te})$
LASSO $\lambda=3 \rightarrow \theta_3(\mathbf{x}, y) \rightarrow p_3(\mathbf{x}_{te}, y_{te})$

$$p(\mathbf{x}_{te}, y_{te}) = \max(p_1(\mathbf{x}_{te}, y_{te}), p_2(\mathbf{x}_{te}, y_{te}), p_3(\mathbf{x}_{te}, y_{te}))$$

↑
 (\mathbf{x}, y)

↑
 $(\mathbf{x}_{te}, y_{te})$

λの選択に使うデータ
評価に使うデータが同じ

上記max操作は
次のようにλの選択
と選択したモデルによる
評価計算に分割可能。

$$\lambda_{te} = \operatorname{argmax}(p_1(\mathbf{x}_{te}, y_{te}), p_2(\mathbf{x}_{te}, y_{te}), p_3(\mathbf{x}_{te}, y_{te}))$$
$$p(\mathbf{x}_{te}, y_{te}) = p_{\lambda_{te}}(\mathbf{x}_{te}, y_{te})$$

アルゴリズムパラメータの選択が 必要なケース

重みの学習

途中評価

評価モデルの選択

LASSO $\lambda=1 \rightarrow \theta_1(\mathbf{x}, y) \rightarrow p_1(\mathbf{x}_{te}, y_{te})$
LASSO $\lambda=2 \rightarrow \theta_2(\mathbf{x}, y) \rightarrow p_2(\mathbf{x}_{te}, y_{te})$
LASSO $\lambda=3 \rightarrow \theta_3(\mathbf{x}, y) \rightarrow p_3(\mathbf{x}_{te}, y_{te})$

$\lambda_{te} = \text{argmax}(p_1(\mathbf{x}_{te}, y_{te}), p_2(\mathbf{x}_{te}, y_{te}), p_3(\mathbf{x}_{te}, y_{te}))$



(\mathbf{x}, y)



$(\mathbf{x}_{te}, y_{te})$



$(\mathbf{x}_{ev}, y_{ev})$

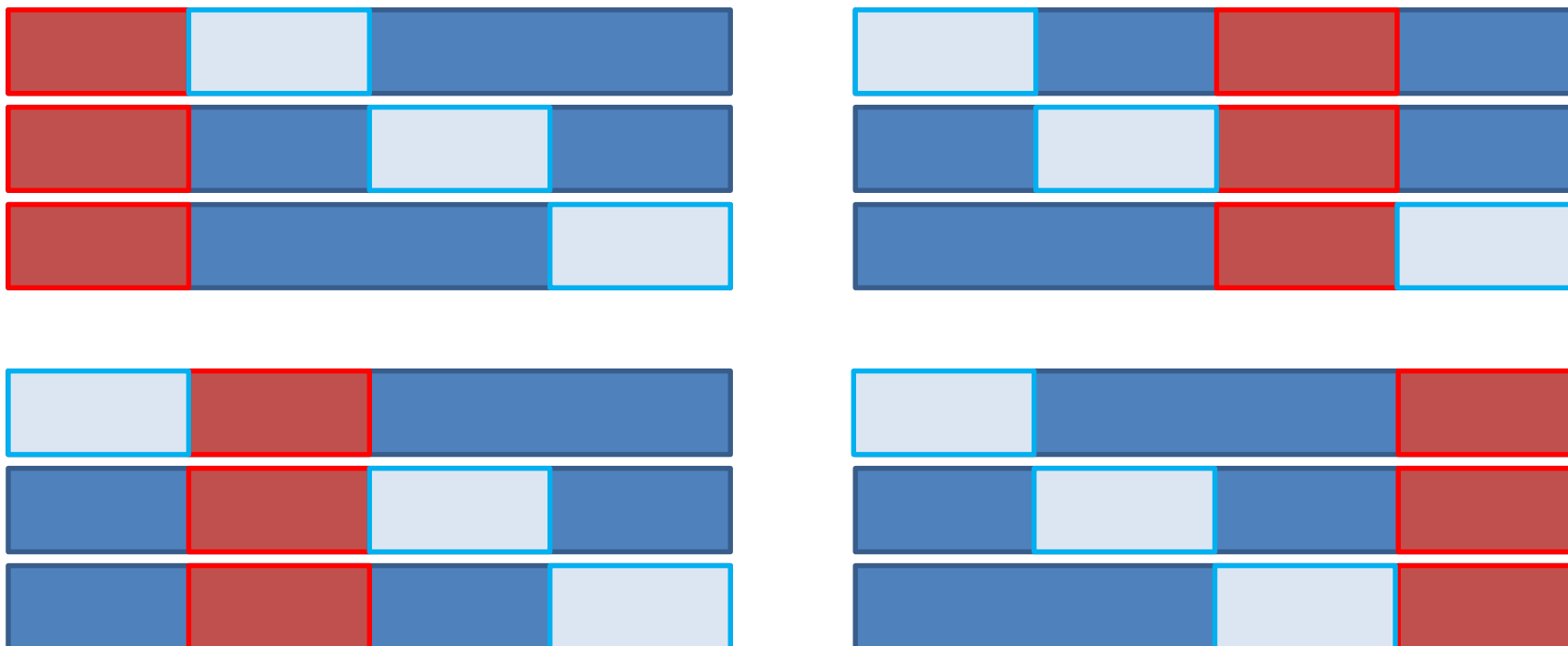
$p_{\lambda_{te}}(\mathbf{x}_{ev}, y_{ev})$

最終評価



予測誤差を評価する方法：発展版

Nested Cross validation



評価

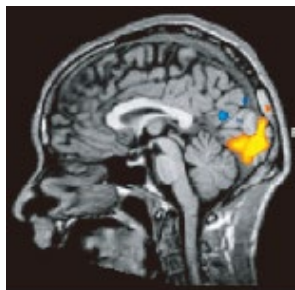
モデル選択

重み学習

3. 情報漏えいにはくれぐれも
気をつけましょう。

ニューロイメージングデータからの病態診断

0. 実験・計測



- イメージングデータ : I



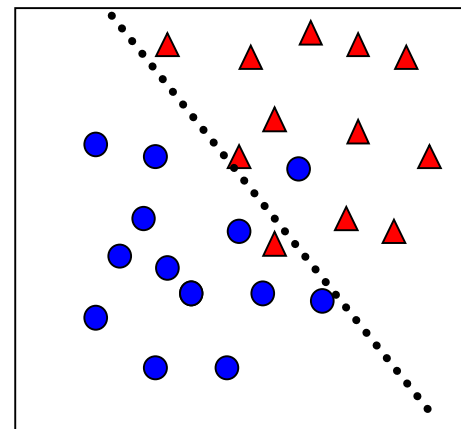
- 病態ラベル : y

1. 特徴量計算



- 画像前処理
- 変数の選択

2. 判別分析



特徴量から病態を
予測するモデル

$$y = f(x)$$

次の手続きに情報漏えいはあるか？

① データ計測

ある課題時の脳活動から、健常者とある疾患患者を判別するために、健常者群・患者群のそれぞれ50人の脳活動をfMRIで計測した。

② 特徴量計算

まず、判別に用いる特徴量を絞るために、

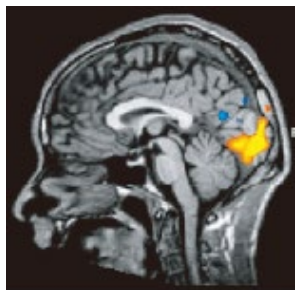
1. 各ボクセルについて、全100人のデータを使って2群間のT検定を行い、
2. $|T| > 1.57$ より大きいボクセルを判別解析用のボクセルとしてスクリーニングした。

③ 予測モデルの学習・評価

次に選ばれたボクセルをBOLD信号を特徴量として、1-subject-out-cross-validation法を用いて、LASSOを学習・評価した。

情報漏えいは特徴量計算も含めた 手続きを考えるとしばしば起こる。

0. 実験・計測



- イメージングデータ : I



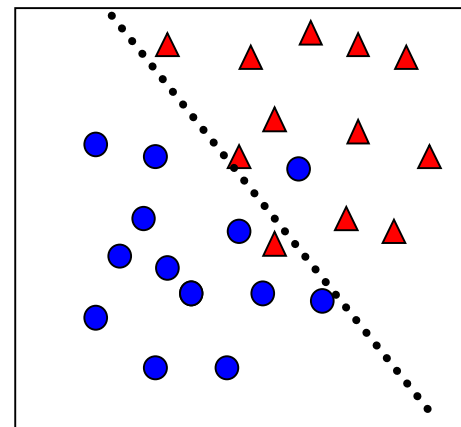
- 病態ラベル : y

1. 特徴量計算



- 画像前処理
- 変数の選択

2. 判別分析

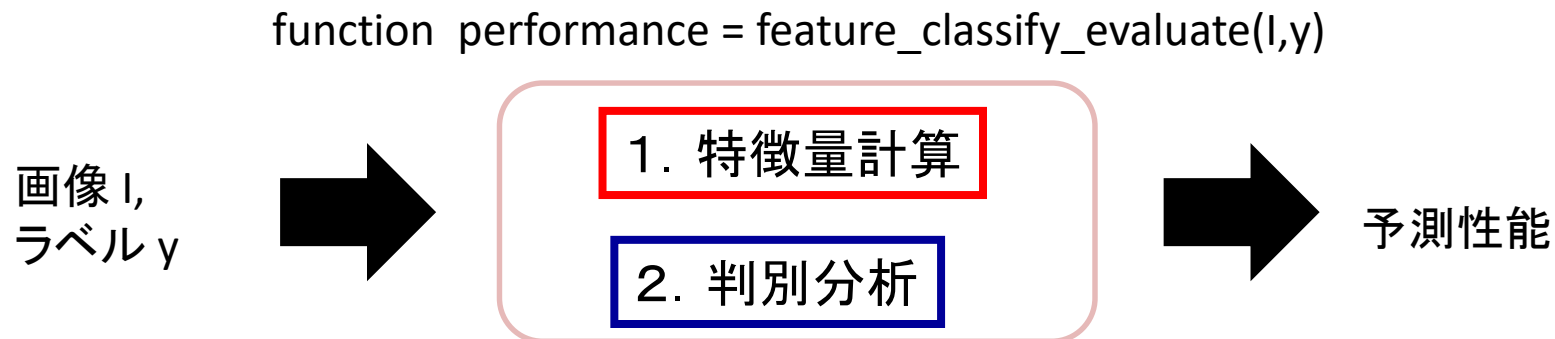


特徴量から病態を
予測するモデル

$$y = f(x)$$

情報漏えいをチェックする方法

1. `function performance = feature_classify_evaluate(I,y)`を作成。
2. `I`と`y`の順番をランダムにシャッフルした`Ishuf`, `yshuf`を作成。
3. `performance = feature_classify_evaluate (Ishuf,yshuf)`の`performance`がチャンスレベルであるかどうかをチェックする。



まとめ

- 予測性能を計算する時は、必ず評価用のデータを独立に用意しましょう。
- 予測モデルを学習するいかなる過程にも、評価用のデータが混入しないように注意しましょう。

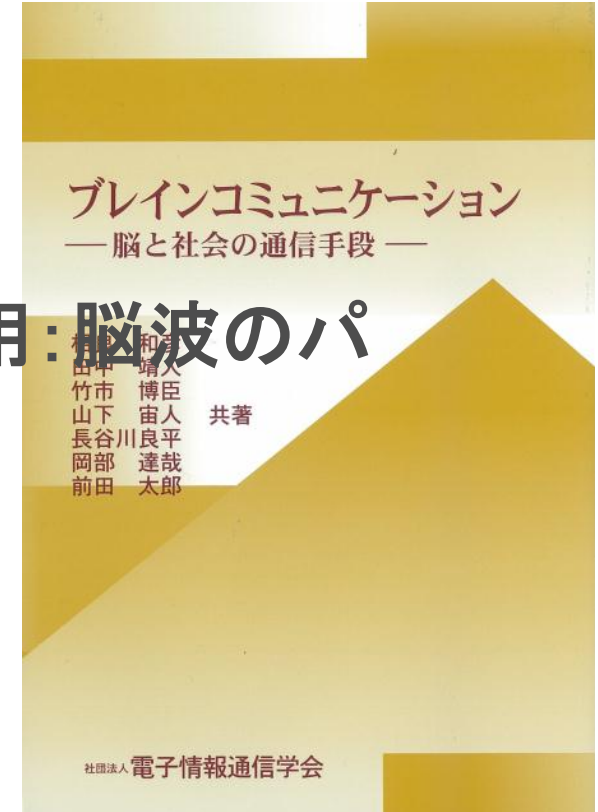
For further study

“Circular analysis in systems neuroscience – the dangers of double dipping”

[Nikolaus Kriegeskorte](#), [W Kyle Simmons](#), [Patrick SF Bellgowan](#), and [Chris I Baker](#), [Nat Neurosci. May 2009; 12\(5\): 535–540.](#) doi: [10.1038/nn.2303](https://doi.org/10.1038/nn.2303)

講義内容

1. 機械学習について
2. モデルの複雑さとオーバフィット
3. 情報漏洩
4. **機械学習のBMIへの応用: 脳波のパターン判別**
5. まとめ



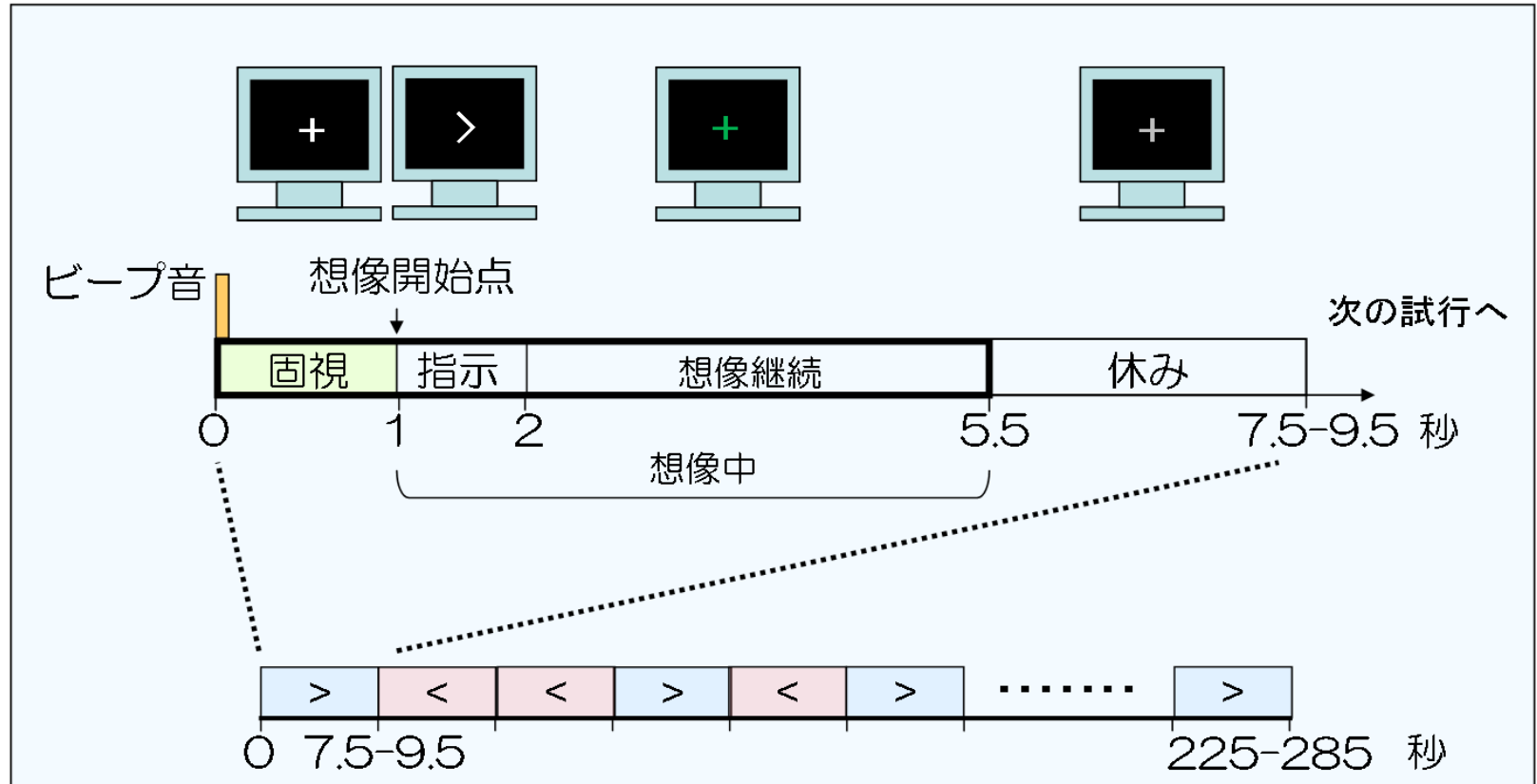
シナリオ

- BMIを作成するために、右手運動想像時と左手運動想像時の脳波データを1試行ごとにパターン分類したい。
- 前処理を行い抽出した特徴量に対してパターン判別を行うため、機械学習法を用いる。
- 先行研究で報告された被験者平均で検出された少数のセンサを用いるケース(低い次元の特徴量)と全64センサを用いるケース(高い次元の特徴量)
- それぞれに対して単純な機械学習法 (Naïve Bayes) と賢い機械学習法 (Sparse Bayes) を適用。
- どのケースが優れているだろうか？

| | パラメータの次元 低 | パラメータの次元 高 |
|---------------------------|---------------|---------------|
| 単純な機械学習法 (Naïve Bayes) | 判別正答率？ | 判別正答率？ |
| 賢い機械学習法 (Sparse Bayes) | 判別正答率？ | 判別正答率？ |

実験：左手右手運動想像課題

実験概要



1ラン = 15試行 × 2条件 = 30 試行

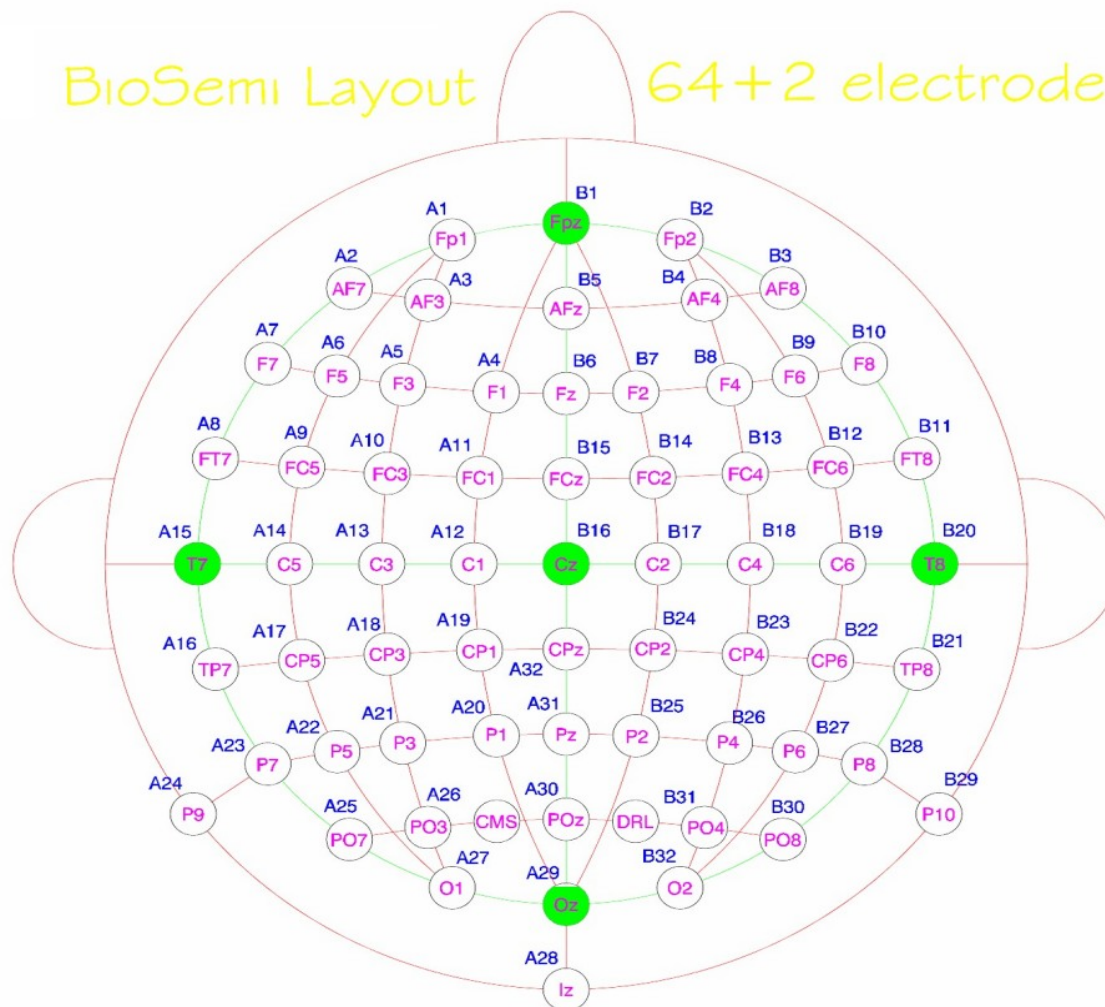
1実験 = 7ラン = 210 試行

計測 : 64ch脳波

EEG : Active two (Biosemi社),
64ch, 全頭, 256Hz

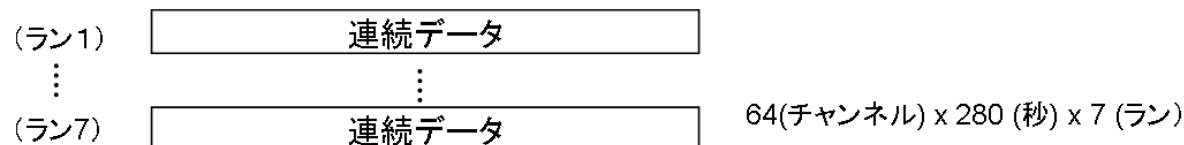
BioSemi Layout

64+2 electrodes



データ解析

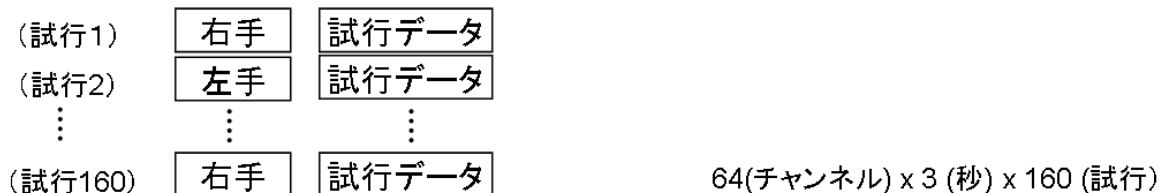
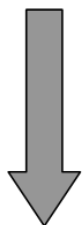
運動想像 課題実験



前処理

1. 基線補正 2. ハイパスフィルタ 3. ローパスフィルタ 4. リリファレンシング

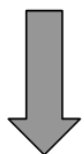
試行データ切り出し
アーティファクト試行の除去



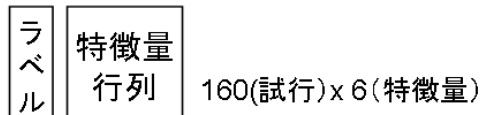
特徴量 計算

パワースペクトル密度推定による9-13Hzと18-25Hzの振動成分のパワーの計算

(特徴量1) Cp3, Cz, Cp4 の3チャンネル (特徴量2) 全64チャンネル



ラベルは
右手=0
左手=1



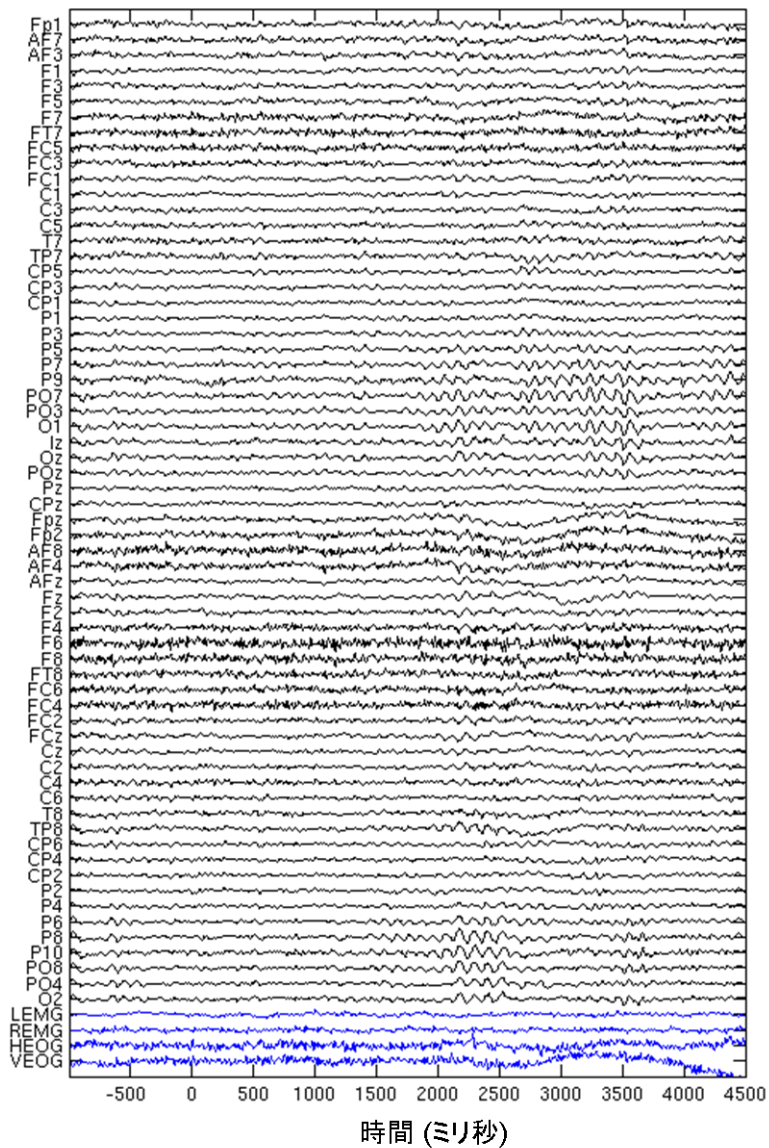
判別

8-fold 交差検定の学習データを用いた判別器の学習
(比較のため以下の3つの判別器を適用: ガウス判別器、SVM、SLR)

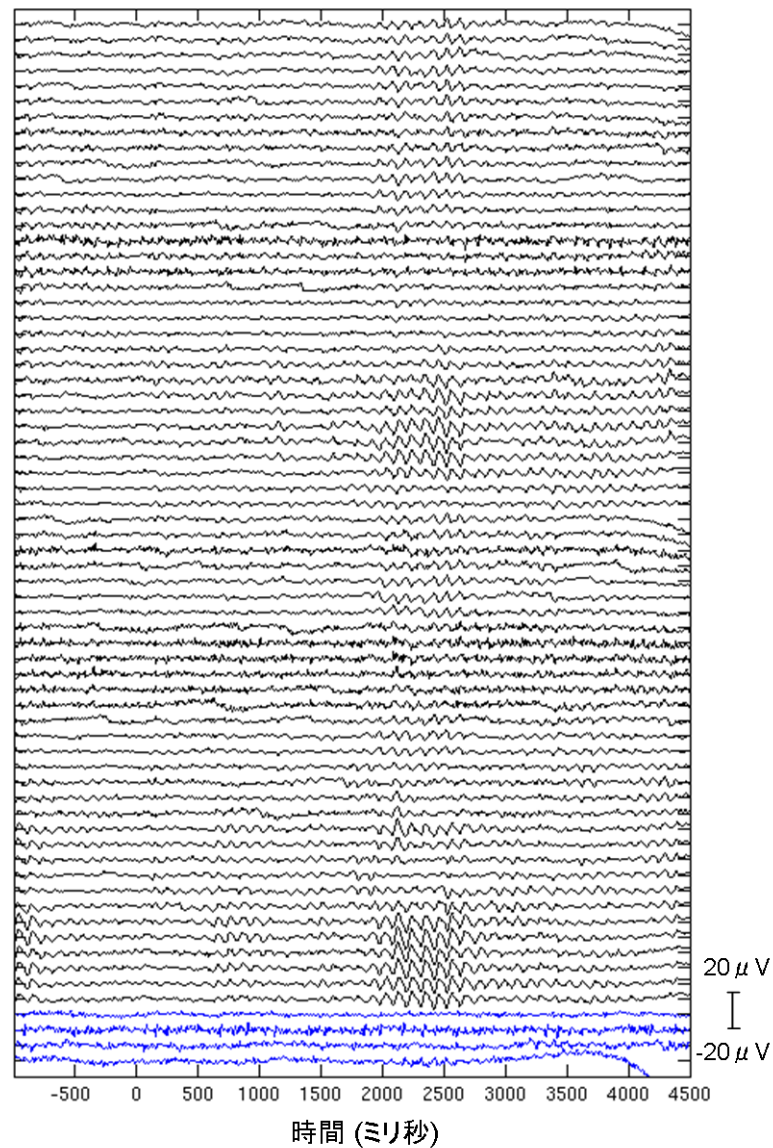
8-fold 交差検定によるテストデータに対する正答率の評価

計測データ例

左手運動想像条件(試行180)

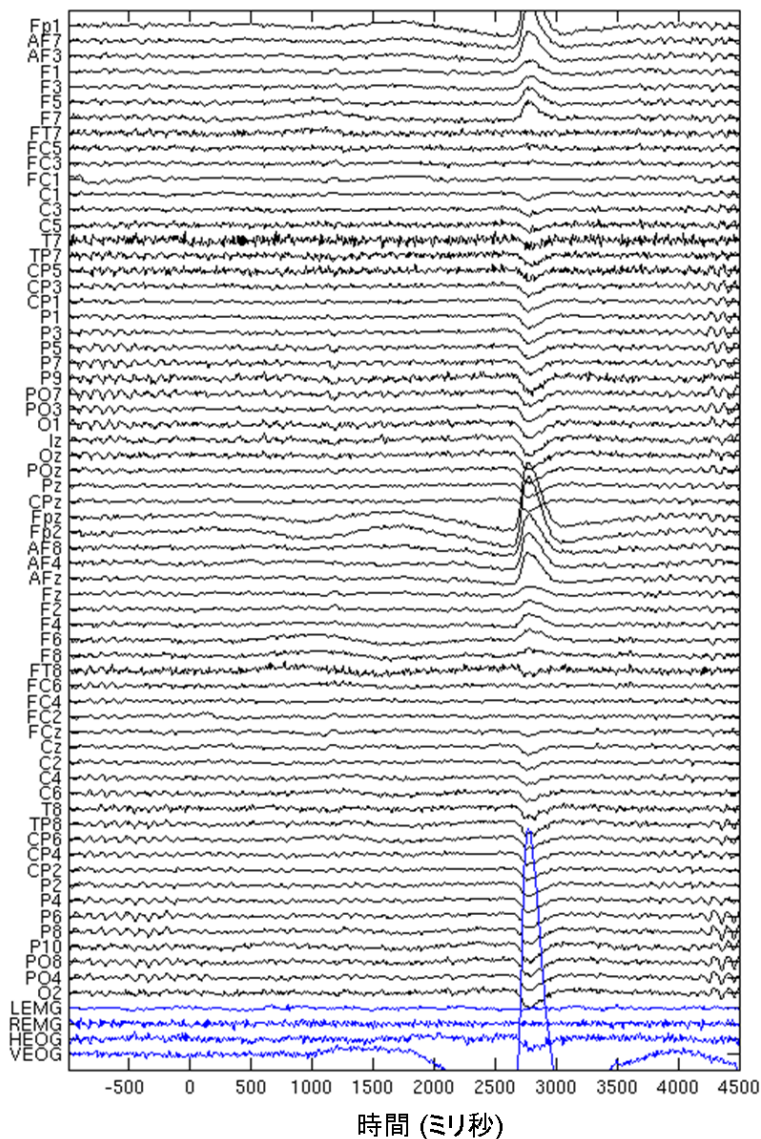


右手運動想像条件(試行156)

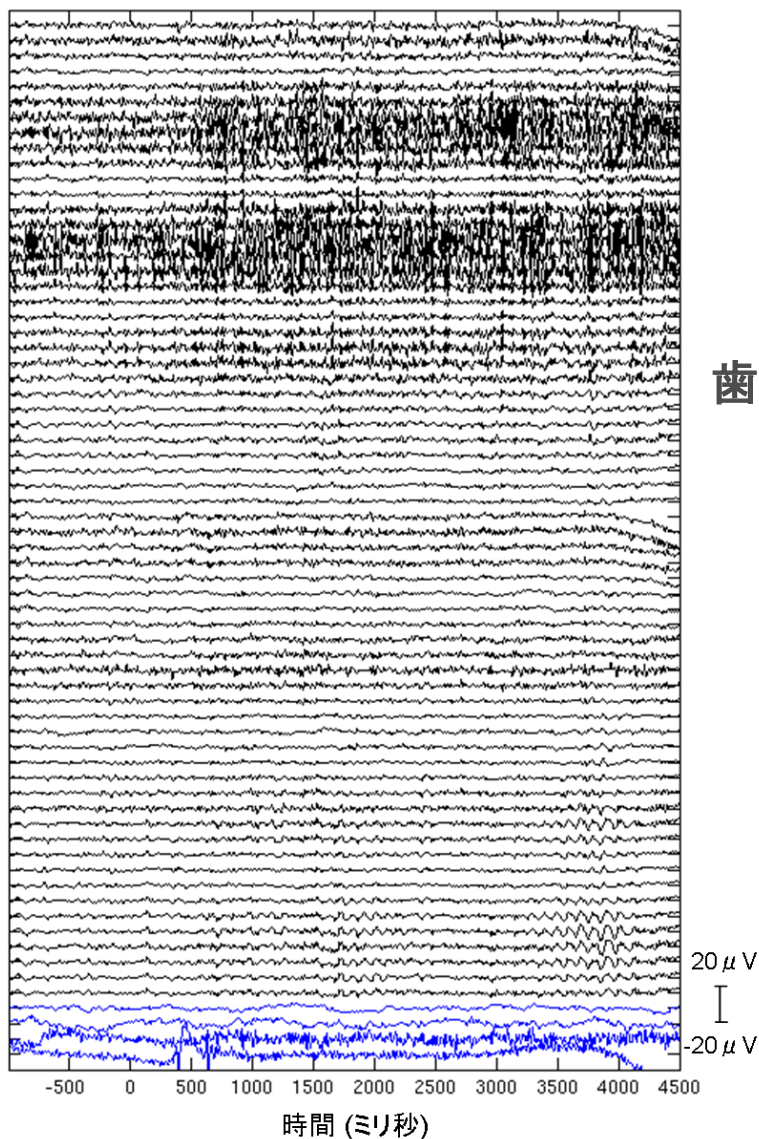


計測データ例：アーチファクト

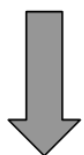
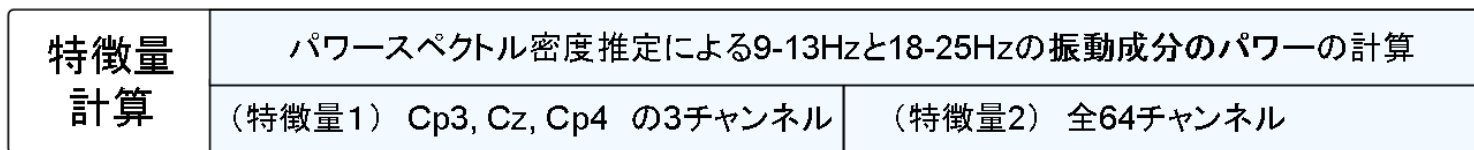
左手運動想像条件(試行166)



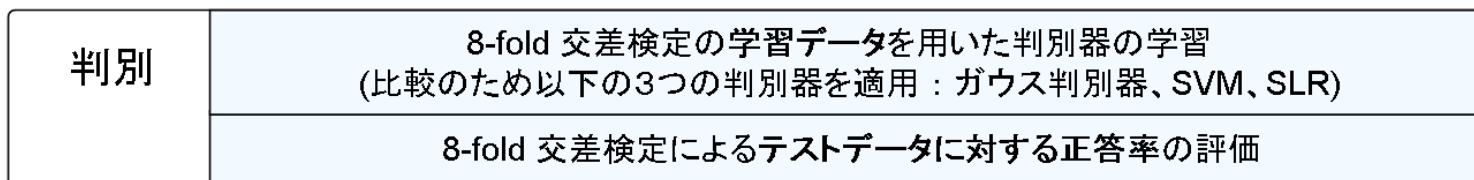
右手運動想像条件(試行4)



データ解析：判別分析

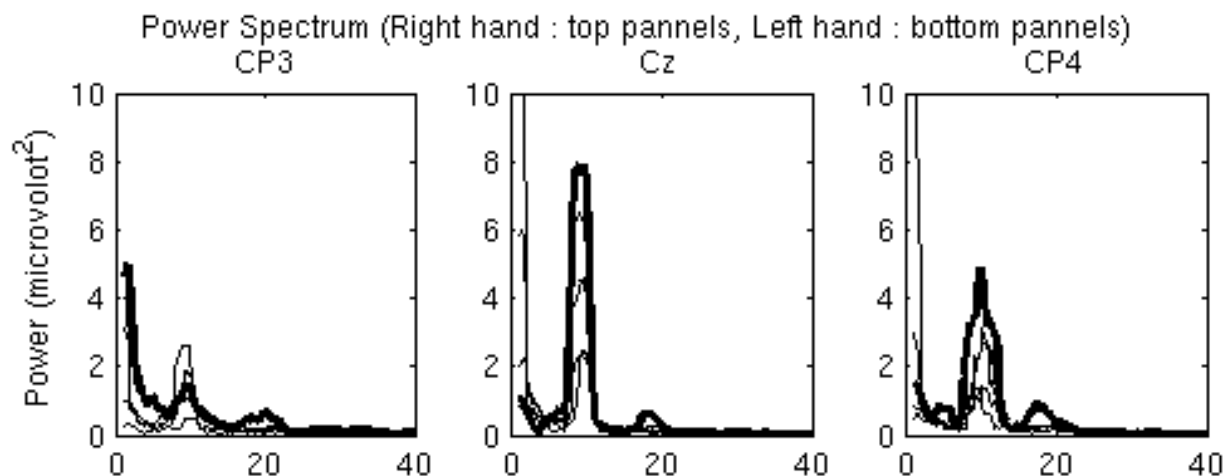


ラベルは
右手=0
左手=1

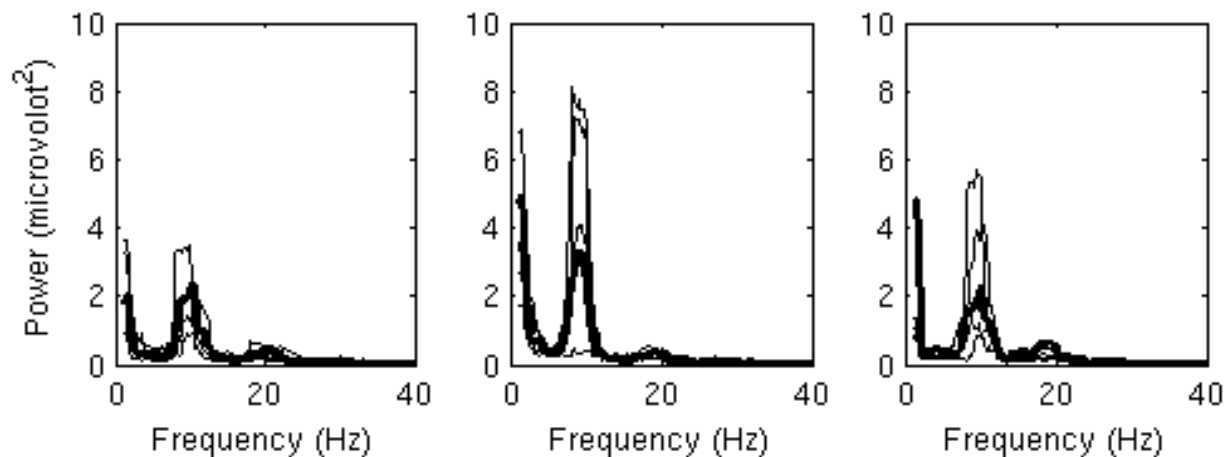


周波数特徴量 (ランダムにサンプルした5試行)

右手
想像



左手
想像



判別結果

単純な判別機
Naïve Bayes

賢い判別機
Sparse Bayes

| | ガウス判別器 | | SLR | |
|-----------------|--------------|--------------|--------------|---------------|
| | 学習 | テスト | 学習 | テスト |
| 特徴量1 (6次元) | 74.9 ±1.6 | 72.7 ±9.5 | 74.1 ±1.6 | 71.1 ±10.4 |
| 特徴量2 (128次元) | 100 ±0.0 | 59.6 ±9.9 | 94.1 ±1.5 | 79.6 ±8.4 |

Model 1 $y = w_1 x_{C3,\alpha} + w_2 x_{C3,\beta} + w_3 x_{Cz,\alpha} + w_4 x_{Cz,\beta} + w_5 x_{C4,\alpha} + w_6 x_{C4,\beta} + c$

Model 2 $y = w_1 x_{F1,\alpha} + w_2 x_{F1,\beta} + w_3 x_{Fz,\alpha} + w_4 x_{Fz,\beta} + \dots + w_{127} x_{O2,\alpha} + w_{128} x_{O2,\beta} + c$

学習した判別器を解釈：スパース正則化によって選ばれた変数

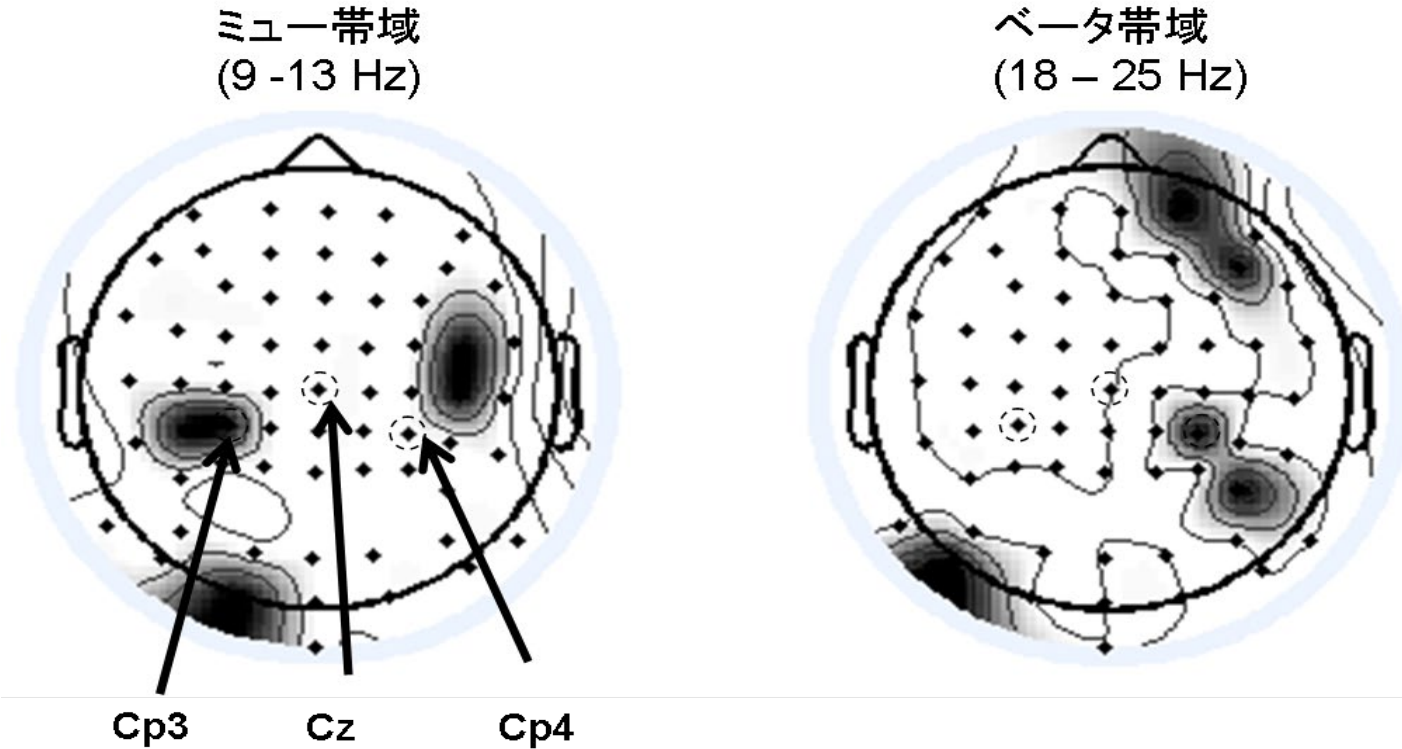


図4.8 SLRで選択された特徴量(チャンネル)を頭皮上にマップした図。ミューリズムで判別に有効なチャンネルを見ると、文献上報告されているCzは選択されておらず、右運動野もCp4の前よりのチャンネルが選択されている。ミュー帯域、ベータ帯域でともに観察される左後ろ側のチャンネルや、ベータ帯域の前側のチャンネルはアーティファクト混入の可能性が考えられる。

講義内容

1. 機械学習について
2. モデルの複雑さとオーバフィット
3. 情報漏洩
4. 機械学習のBMIへの応用:脳波のパターン判別
- 5. まとめ**

- モデルの複雑さとオーバフィット・アンダーフィット。
 - オッカムの剃刀の原理
 - モデル選択基準
 - 正則化法

- 情報漏洩の問題
 - クロスバリデーション法
 - Nestedクロスバリデーション法
 - 特徴量選択やハイパーパラメータの選択時は特に注意が必要

有効な変数を自動選択する
スパース制約を用いた判別手法

ATR 脳情報研究所
山下 宙人



12
45



Contents lists available at [ScienceDirect](#)

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns

Okito Yamashita ^{a,*}, Masa-aki Sato ^a, Taku Yoshioka ^{a,b}, Frank Tong ^c, Yukiyasu Kamitani ^{a,b}

^a *ATR Computational Neuroscience Laboratories, Japan*

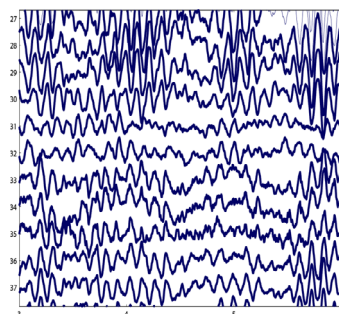
^b *National Institute of Information Technology, Japan*

^c *Vanderbilt University, Psychology Department, USA*

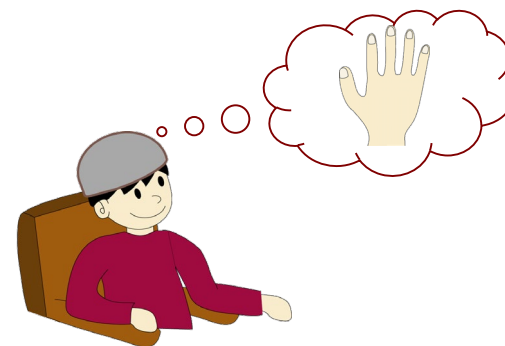
脳活動データの特徴

脳活動データ : Z

認知状態 : Y



$$Y = f(Z)$$



- 個人差が大きい
- 候補となる変数が膨大（時間、空間、周波数）
→ “大規模な変数”の選択問題
- 学習のためのデータ数が少ない
→ 学習データに偏ったパラメータ推定（過学習）

スパースロジスティック判別器 (Sparse Logistic Regression)

ロジスティック回帰モデルとARD事前分布という確率モデルを用いたスパース線形判別器

高次元の特徴量

判別器によって自動的に抽出された低次元の特徴量



学習データの判別を良くする
(尤度：ロジスティック回帰モデル)
+
少ないパラメータの数で説明する
(事前分布：ARD事前分布)



スパースロジスティック判別器

- ロジスティック回帰モデルをベイズ推定に拡張したもの
- スパース事前分布として知られる **Automatic Relevance Determination prior (ARD)** (Mackay 1994, Neal 1996)を適用

ロジスティック回帰モデル

判別モデル

given $(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)$,
 $t_i \in \{0, 1\}$: input , \mathbf{x}_i : output

$$\text{Likelihood} : P(\mathbf{t} | \mathbf{X}; \Theta) = \prod_{i=1}^N p_i^{t_i} (1 - p_i)^{1-t_i}$$

where $p_i = 1 / (1 + \exp(-\mathbf{w}^t \mathbf{x}_i))$,

Prior : $P(\theta_k | \alpha_k) \propto N(0, \alpha_k^{-1})$

Hyper Prior: $P(\alpha_k) \propto \alpha_k^{-1}$ (non-informative)

ARD事前分布

推定アルゴリズム

[Tipping, 2001]

W-step

$$[\hat{\mathbf{w}}] = \arg \max_{\mathbf{w}} \left[\sum_{i=1}^N \{t_i \log p_i + (1 - t_i) \log(1 - p_i)\} - \frac{1}{2} \theta' \langle A \rangle_{\mathbf{a}} \theta \right]$$

A-step

$i = 1, 2, \dots, D$

$$\alpha_i = (1 - \alpha_i S_{ii}) / \hat{\theta}_i^2$$

where $S = (X^t B X + A)^{-1}$

$$(B)_{ii} = p_i(1 - p_i)$$

$$(A)_{ii} = \alpha_i$$

$\langle A \rangle_{\mathbf{a}}$: expectation of A
with respect to $\alpha_1, \dots, \alpha_D$

L1 norm 正則化とスパース化

最小二乗法

$$E(\mathbf{w}) = \underbrace{\|y - \mathbf{w}^t \mathbf{x}\|^2}_{\text{フィッティング}}$$

L2-norm正則化

$$E(\mathbf{w}) = \underbrace{\|y - \mathbf{w}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\|\mathbf{w}\|^2}_{\text{制約}}$$

L1-norm正則化

$$E(\mathbf{w}) = \|y - \mathbf{w}^t \mathbf{x}\|^2 + \lambda \sum_i |w_i|$$

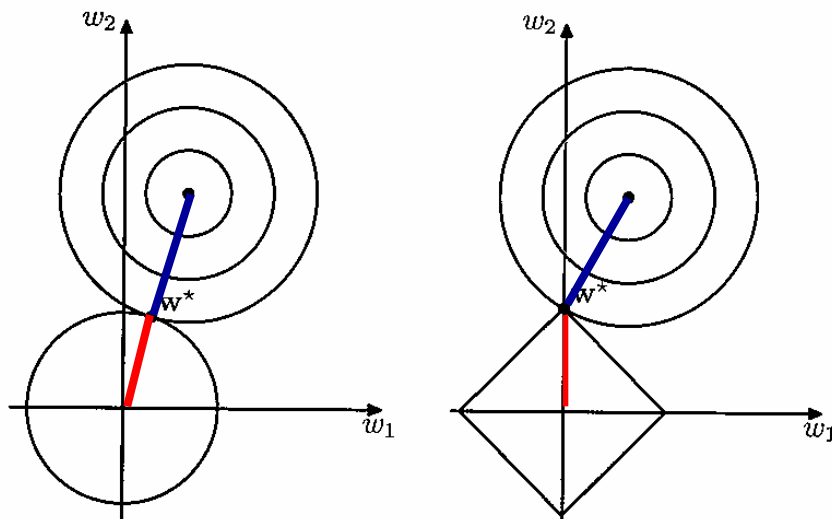
L1 norm 正則化とスパース化

L2-norm正則化

$$E(\mathbf{w}) = \underbrace{\|y - \mathbf{w}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\|\mathbf{w}\|^2}_{\text{制約}}$$

L1-norm正則化

$$E(\mathbf{w}) = \underbrace{\|y - \mathbf{w}^t \mathbf{x}\|^2}_{\text{フィッティング}} + \lambda \underbrace{\sum |w_i|}_{\text{制約}}$$

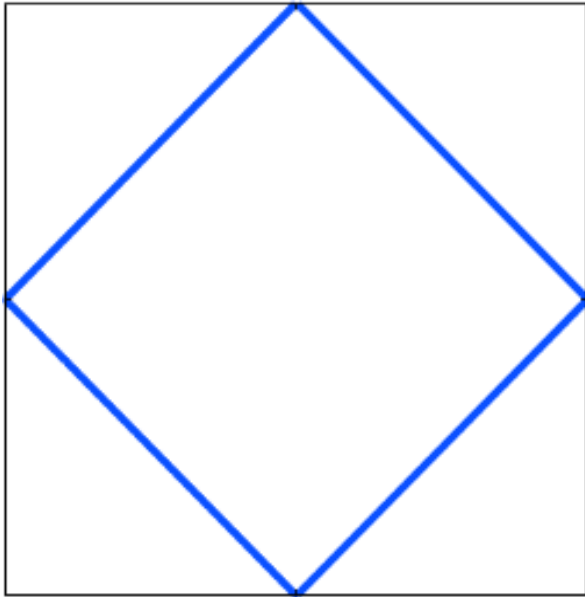


L2-norm

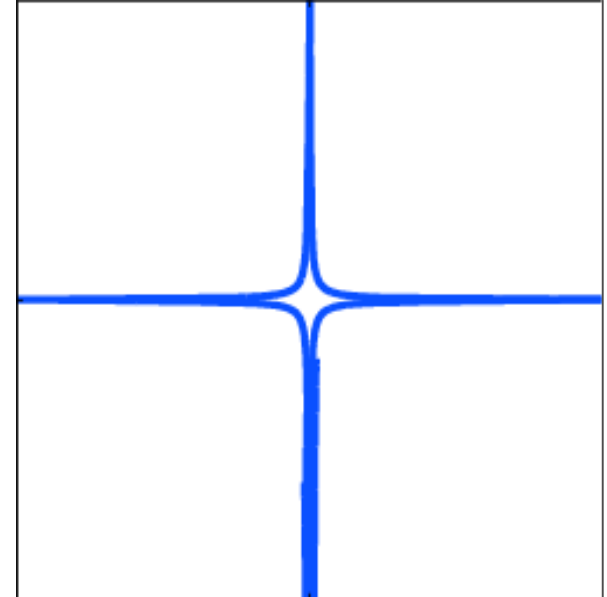
L1-norm

ARDの正則化項

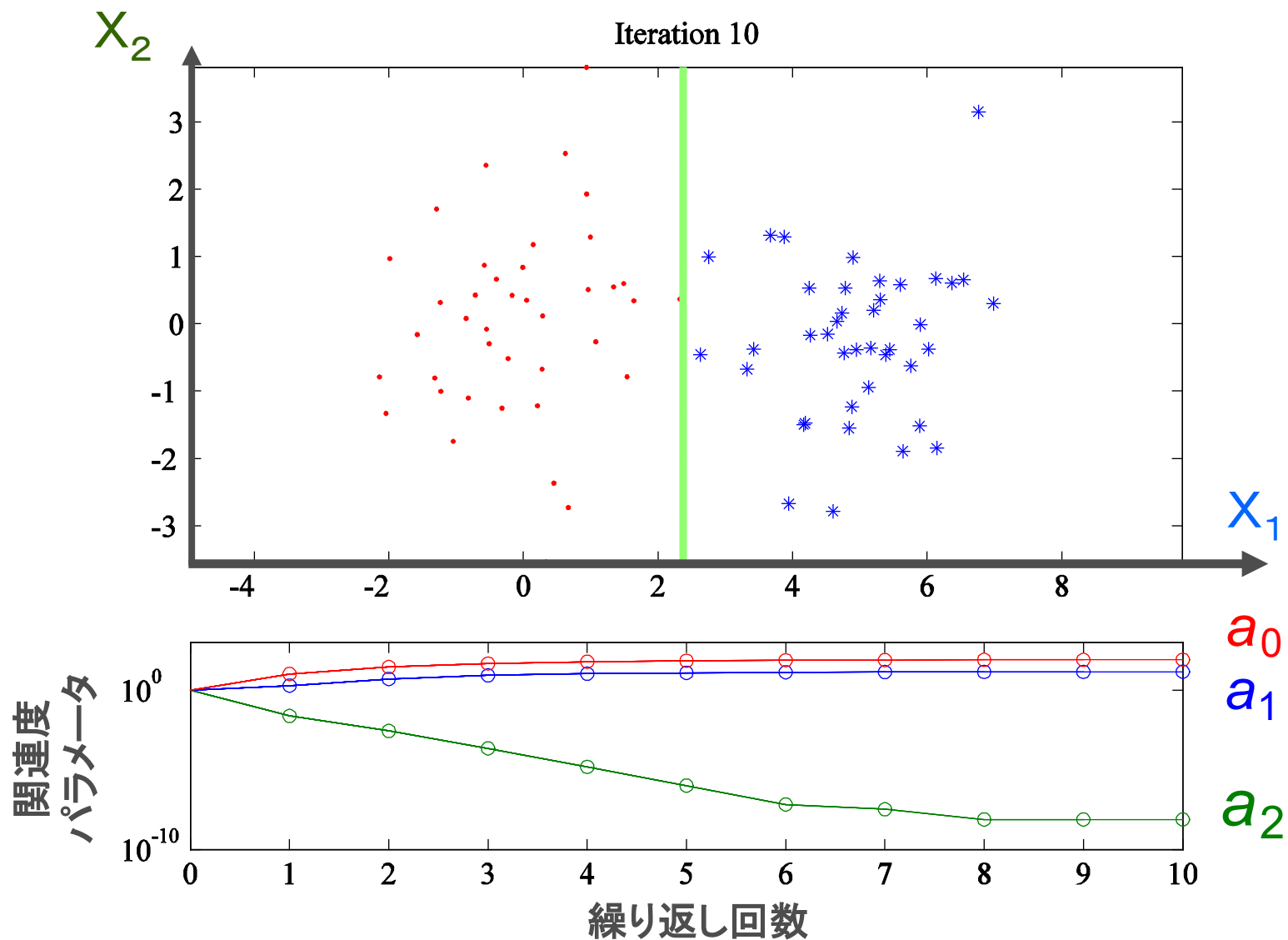
L1-norm
Laplace



ARD



線形判別境界: $w_1x_1 + w_2x_2 + w_0 = 0$ の例



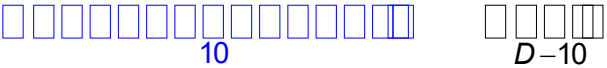
シミュレーションデータ1 - ノイズ次元に対する頑健性 -

データ生成

Gauss分布から生成

$$\mu_1 = [0.1 \ 0.2 \ \dots \ 0.9 \ 1.0 \ 0 \dots \ 0]$$

$$\mu_2 = [0 \ 0 \ \dots \ 0 \ 0 \ 0 \dots \ 0]$$



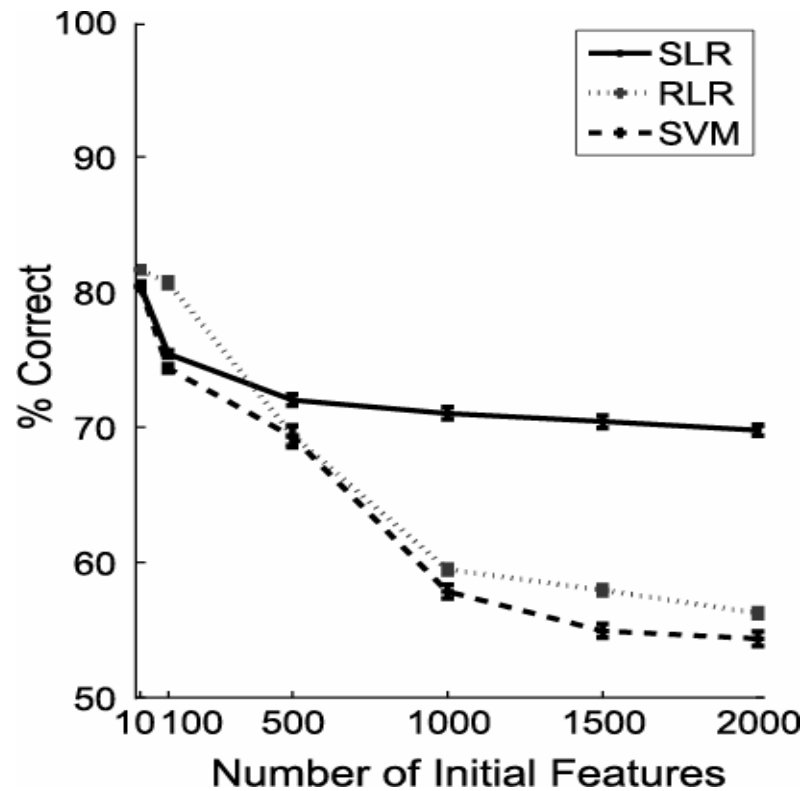
10 D-10

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$

解析

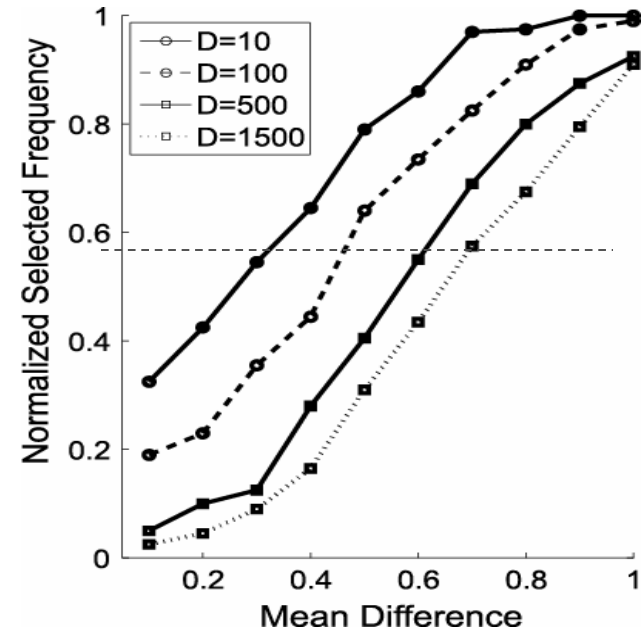
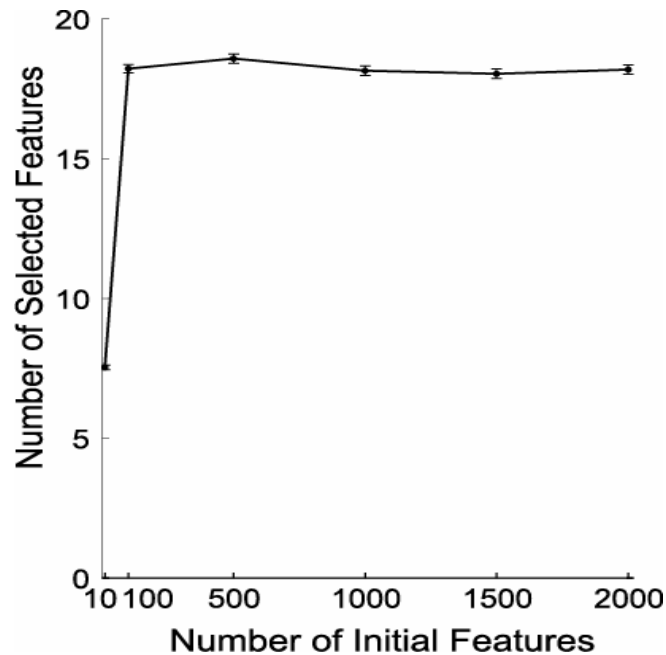
$D = 10, 100, 500, 1000, 1500, 2000$ について判別率を評価

SLRはノイズ次元が大きくても判別率の低下は緩やかである



libsvm2.8.2 :
As a trade-off parameter,
the defaults value (=1) was used.

SLRは少数の有効な次元を選択する。
高い判別性を持つ次元から選択される。



$D=10$ の時に過度なスパース化が起こっている。

データ生成

Gauss分布から生成

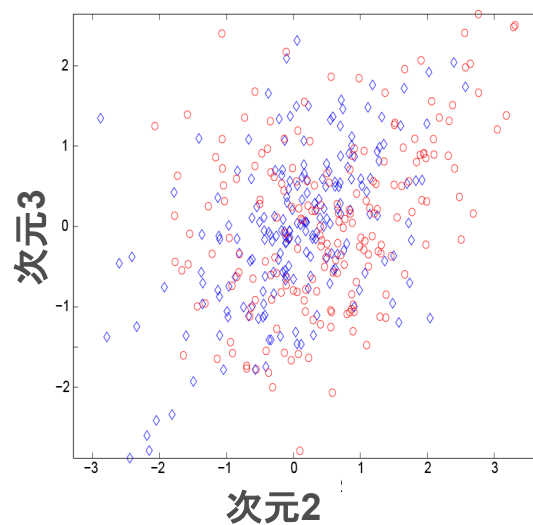
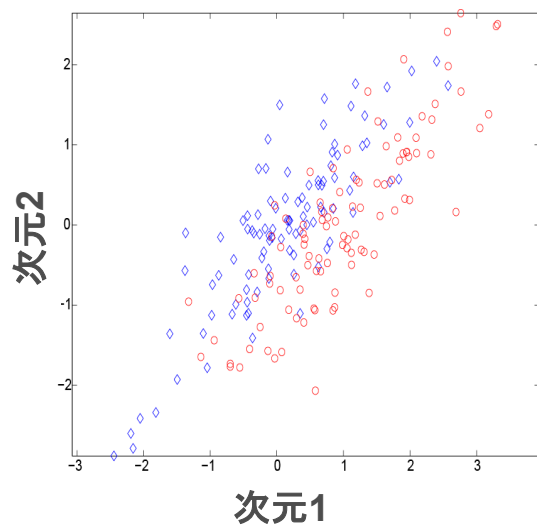
$$\mu_1 = [1 \ 0 \ 0 \ \dots \ 0]$$

$$\mu_2 = [0 \ 0 \ 0 \ \dots \ 0]$$

□ □
2

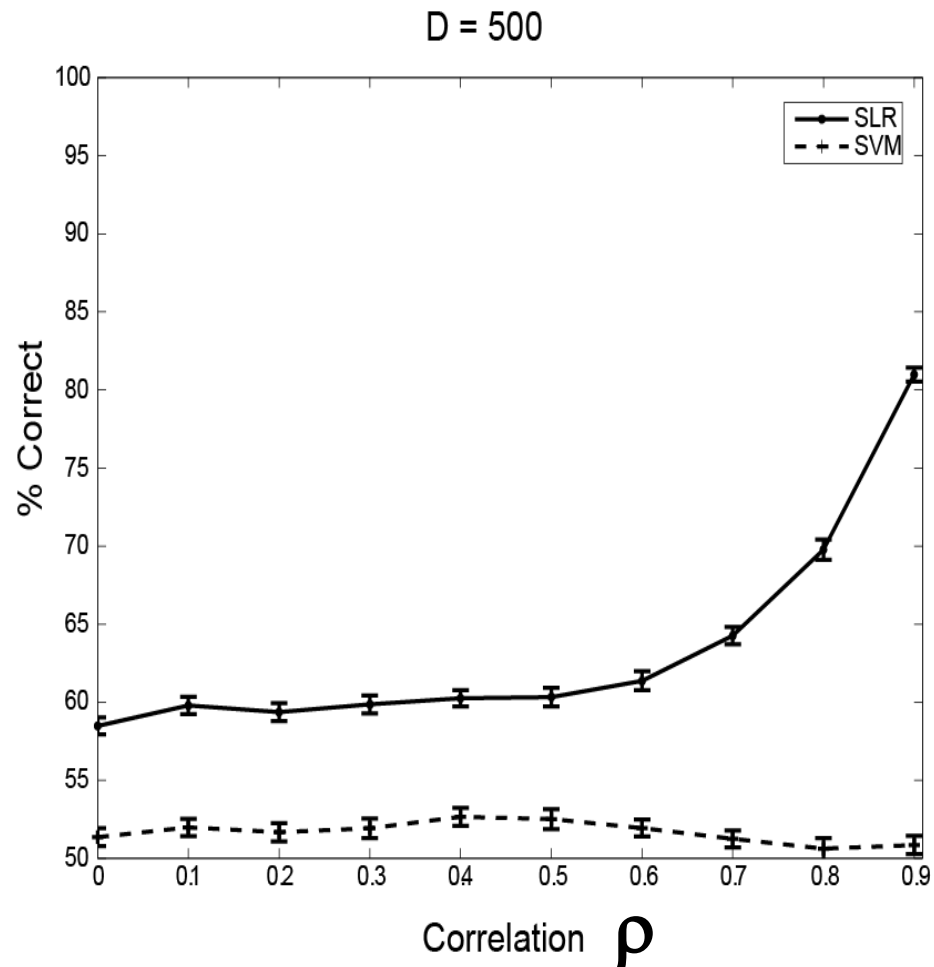
□ □ □ □
498

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & & \\ 0 & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$



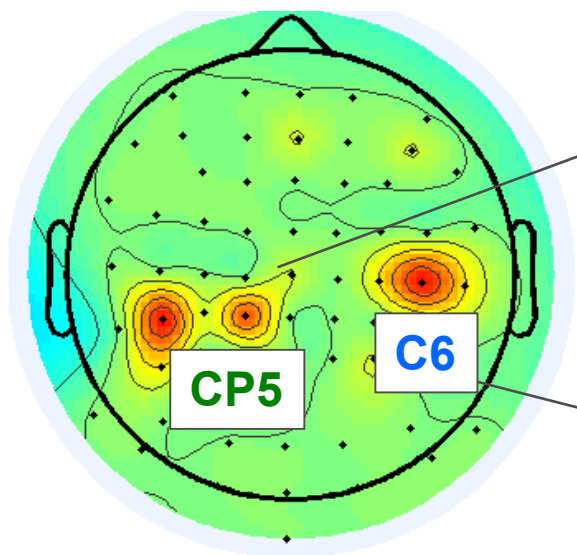
SLRは有効な相関構造を持つ

次元を選択することによって正答率を向上させる。

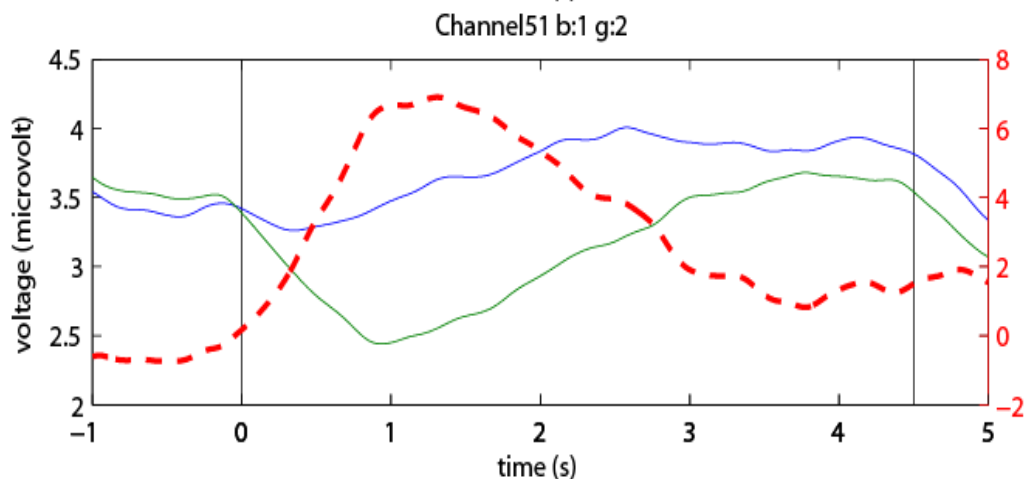
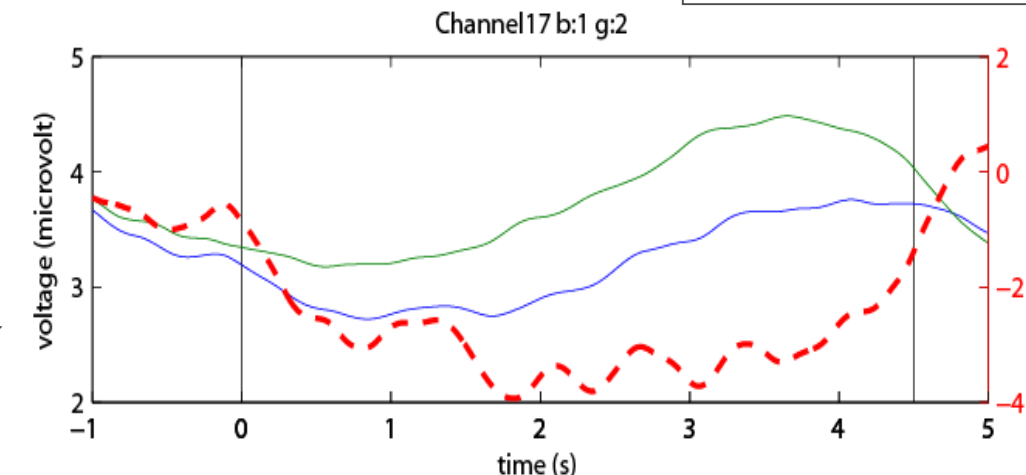
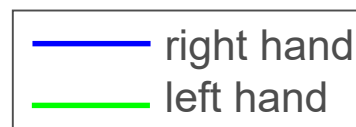


実データ解析 - 左手右手運動想像課題 -

スパース推定で
選ばれるチャンネル

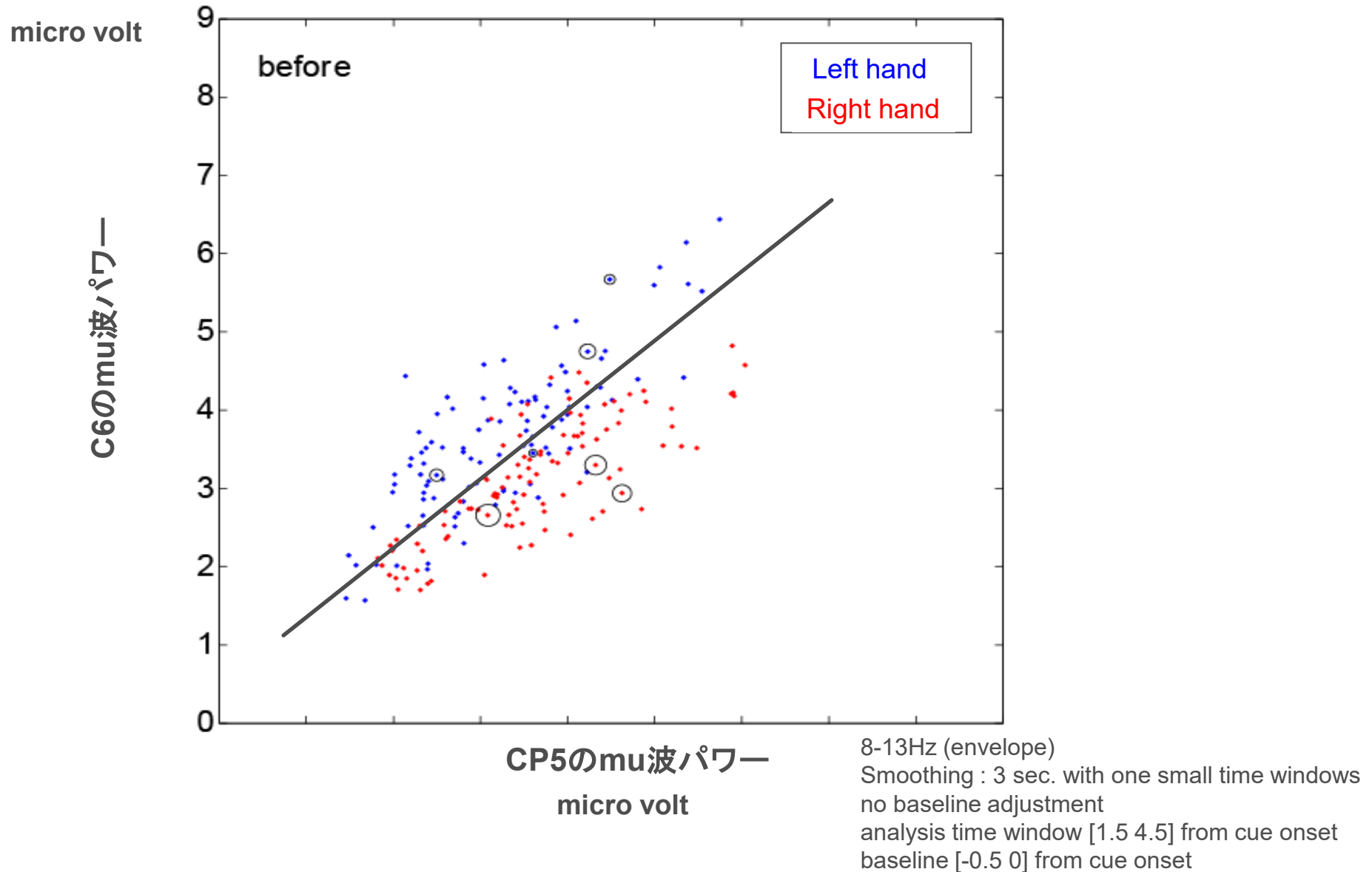


試行平均波形
(mu波のパワー)

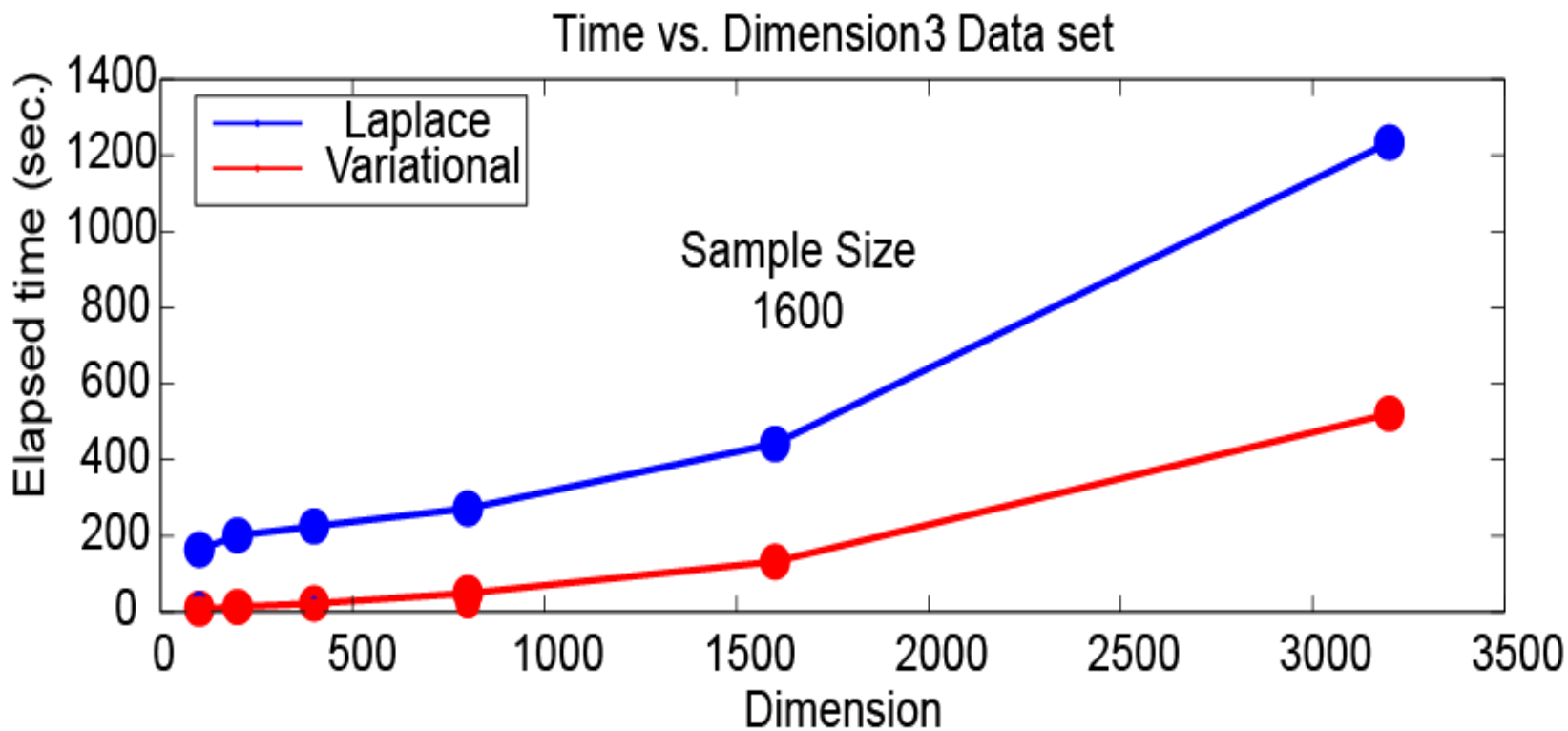


8-13Hz (envelope)
Smoothing : 1 sec.
no baseline adjustment

2つのチャンネルデータの相関



計算時間



1600 x 3200 の問題で10分未満
1600 x 400 の問題で数秒

CPU : 2.66GHz Xeon(R)
Memory : 4GB
MALAB version : 7.0.1

SLRの性質 (データ解析からわかること)

+ノイズとなる次元を削ることによって性能の低下を回避

+平均の違いだけではなく相関も考慮

-弱い判別能力を持つ次元を削ることによる性能の低下

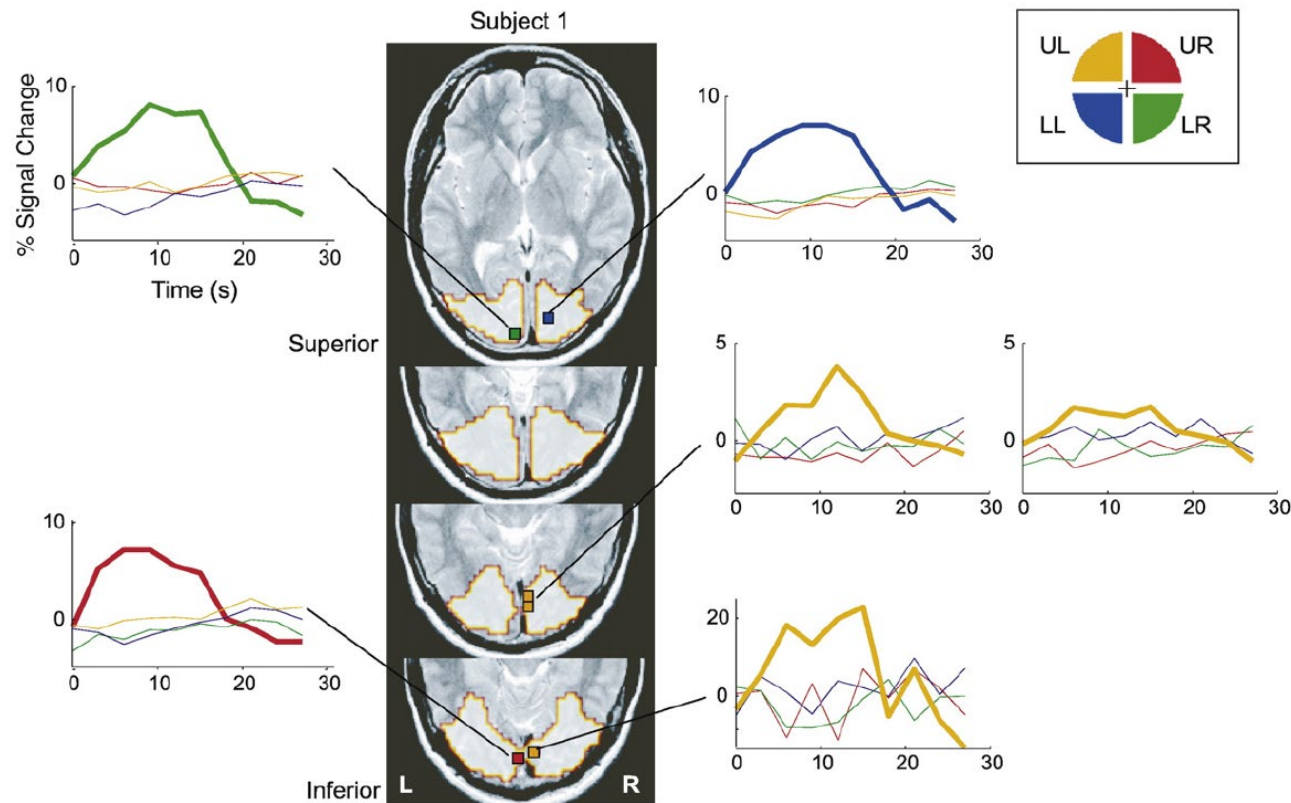
提案手法 の長所

- スパース化による過学習の回避。
- 選択された特徴量による結果の解釈性の向上。
- アルゴリズムにチューニングするパラメータが無い。

初心者にも簡単に使用可能
実用性の高いアルゴリズム

結果解釈時に気をつけること

- 生き残る特徴量の数はサンプル数に依存する。
- 生き残った特徴量はデコーディングするのに十分な特徴量である。エンコーディングの意味で重要な特徴全体とは必ずしも一致しない。

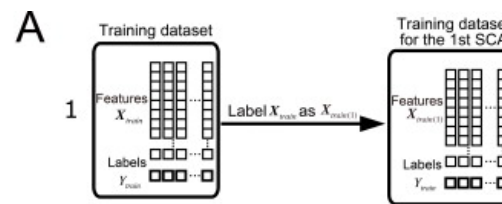


四半視野刺激を
判別するボクセル

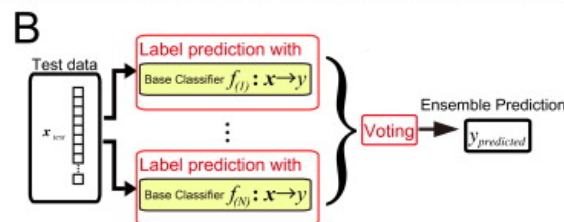
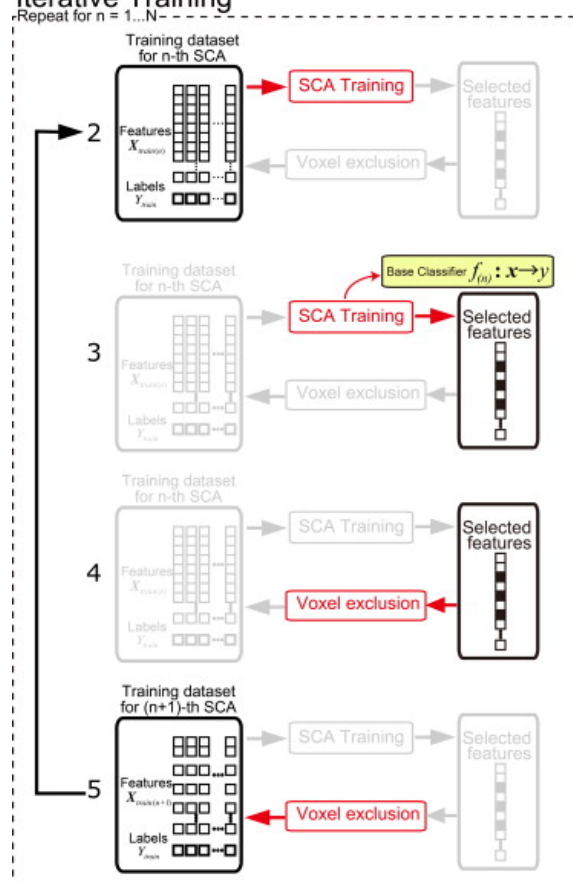
よくある質問

Q. スパースになりすぎるのですがどうしたら良いですか？

A. SLRではスパースの度合いはコントロールできません。CiNET 廣江君が提案した iterative SLR を使いましょう。繰り返し回数はパラメータとなり、設定する必要があります。



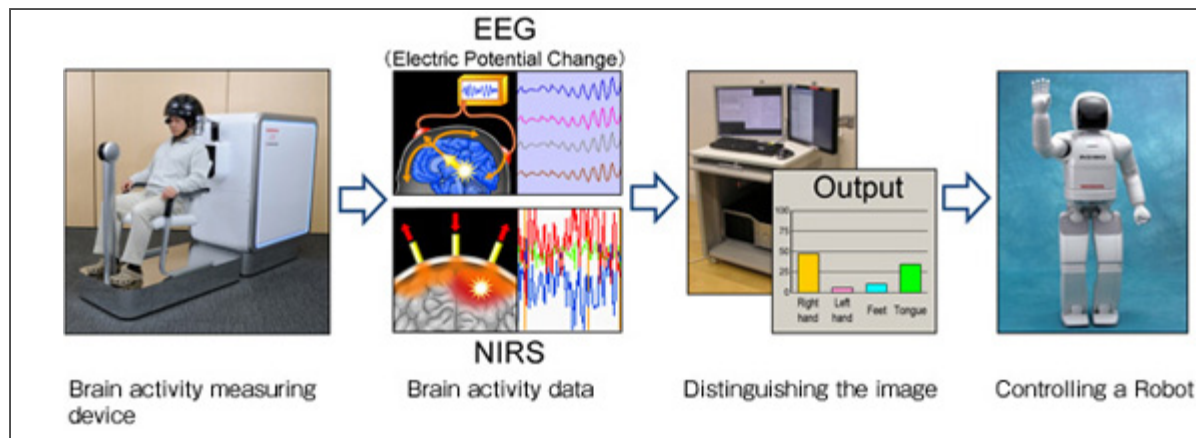
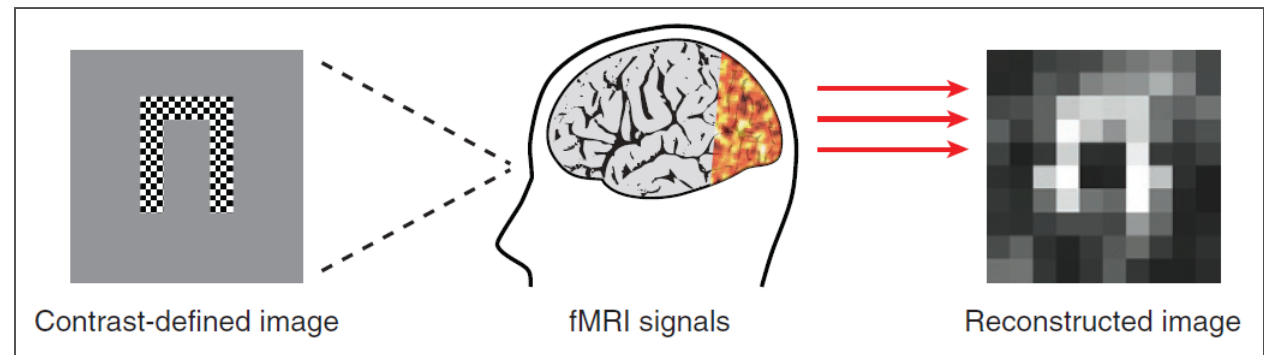
Iterative Training



SLRの適用例

脳活動からの 視覚画像再構成

Miyawaki et.al 2008 *Neuron*



EEG-NIRS BMI

Honda-ATR-Shimazu
による協同研究

MATLAB toolbox が利用可能

スパース推定ライブラリ(MATLAB版)

近年、神経科学の方法、応用の両面において、脳活動から被験者の認知状態や受ける刺激を予測する手法の研究の重要性が増しています。ATR脳情報研究所では、ベイズ学習理論の分野で発展を遂げている“スパース推定”の原理を用いて、関連の無い脳活動を自動的に省きながら予測モデルを構築するアルゴリズムの開発を行い、様々な問題において成果を挙げています。スパース推定アルゴリズムは、

1. 少数サンプル、高次元モデル(数千以上)でもパラメータ推定可能
2. 過学習の回避
3. 結果の可解釈性の向上

などのメリットを持っています。また、アルゴリズムのパラメータのチューニングは基本的に不要なので、幅広いデータにすぐに適用することができます。

本ライブラリでは、我々のグループ及び共同研究者によって開発された、スパース推定のための"MATLAB"ツールボックスを提供します。本ライブラリは、

- 3つのスパース回帰モデルのツールボックス(予測変量が連続値をとる場合)
- 1つのスパース判別モデルのツールボックス(予測変量がカテゴリカルな値をとる場合)

と4つの独立なツールボックスからなります。テスト用のデータも用意されていますので、まずは試用してみてください。

https://bicr.atr.jp//cbi/sparse_estimation/index_j.html

参考文献

David J.C. MacKay(1992), **Bayesian Interpolation**, Neural Computation.

Neal, R. M. (1994). **Priors for infinite networks**, Technical Report In preparation, Univ. of Toronto.

Michael E. Tipping. (2001), **Sparse bayesian learning and the relevance vector machine**, *Journal of Machine Learning Research*.

Wolpaw et.al. (2002), **Brain-computer interfaces for communication and control**. Clin Neurophysiol., pp767-91

Kamitani Y, Tong F (2005), **Decoding the visual and subjective contents of the human brain**. Nat Neurosci., pp679-85

Miyawaki Y. et.al.(2008), **Visual image reconstruction from human brain activity using a combination of multiscale local image decoders**. Neuron, pp915-29

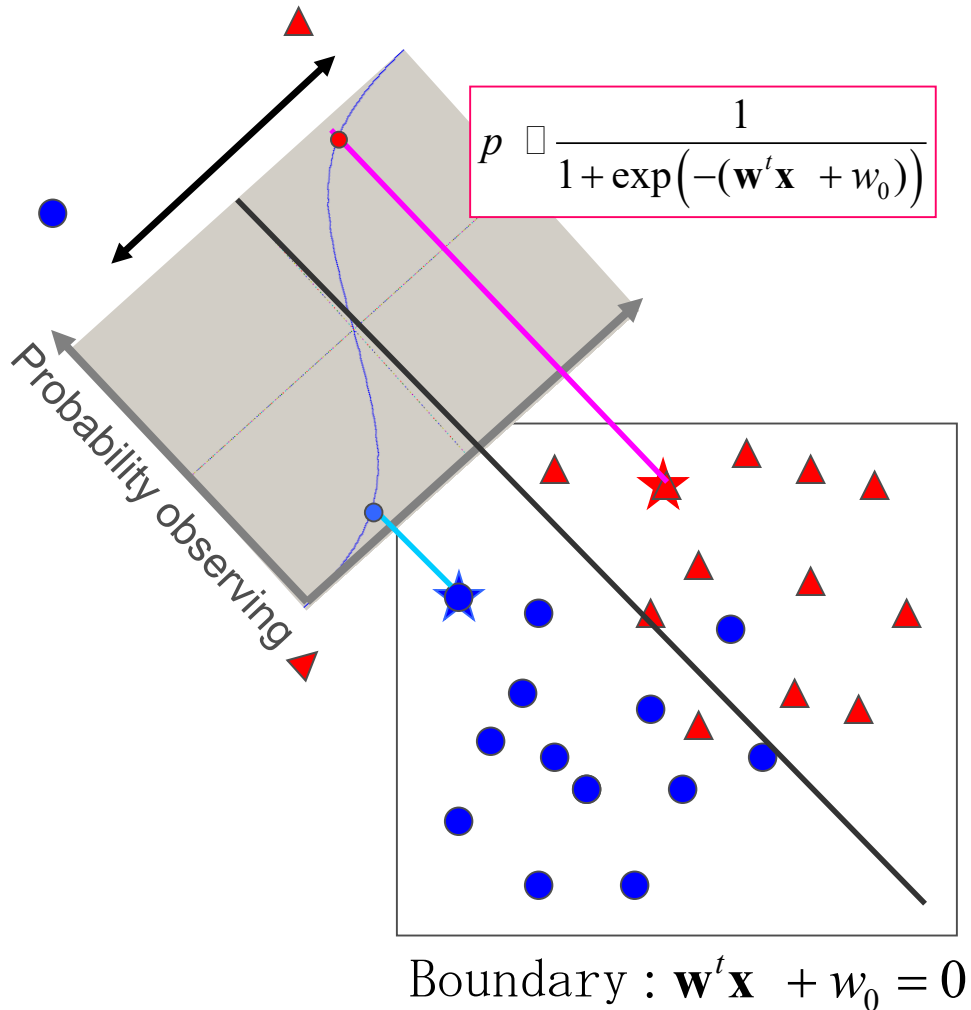
Efron B. et.al (2004), **LARS**

Tibshirani R. (1996), **Regression Shrinkage and Selection via Lasso**, JRSSB, 267-288

アルゴリズム詳細

(Tipping 2001を参照)

- Logistic Regression Model



サンプル: $\{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$

特徴量ベクトル: $\mathbf{x}_i \in \mathbb{R}^D$

カテゴリラベル: $t_i \in \{0, 1\}$

尤度関数: $P(\mathbf{t} | \mathbf{X}; \mathbf{W}) = \prod_{i=1}^N p_i^{t_i} (1 - p_i)^{1-t_i}$

パラメータ: $\mathbf{W} = \{\mathbf{w}, w_0\}$

最尤法
(**Iterative** reweighted **least**
squares method)

- SLRモデル

$$\text{likelihood} : P(\mathbf{t} | \mathbf{X}, \mathbf{W}) = \prod_{i=1}^N p_i^{t_i} (1 - p_i)^{1-t_i} \quad p_i \square \frac{1}{1 + \exp(-(\mathbf{w}'\mathbf{x}_i + w_0))}$$

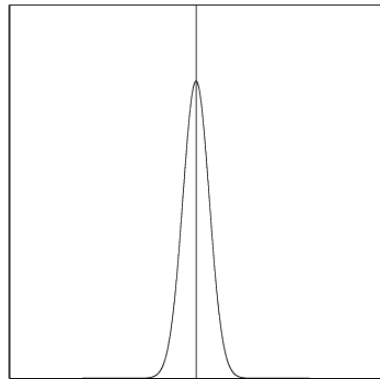
$$\text{Prior} : P(w_k | \alpha_k) \square N(0, \alpha_k^{-1}) \quad k = 1, 2, \dots, D$$

$$\text{Hyper Prior: } P_0(\alpha_k) = \alpha_k^{-1} \quad k = 1, 2, \dots, D$$

ARD priors

α_k^{-1} : small

prior
probability
of w_k

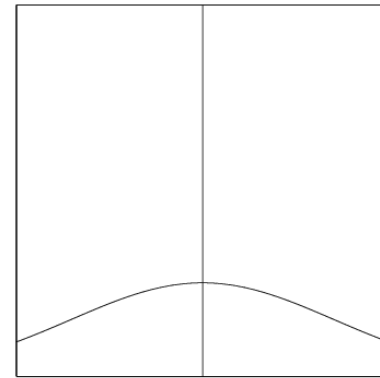


0

w

$w_k = 0$ の周りに大きな事前確率

α_k^{-1} : large



0

w

広い範囲の w_k をとりうる

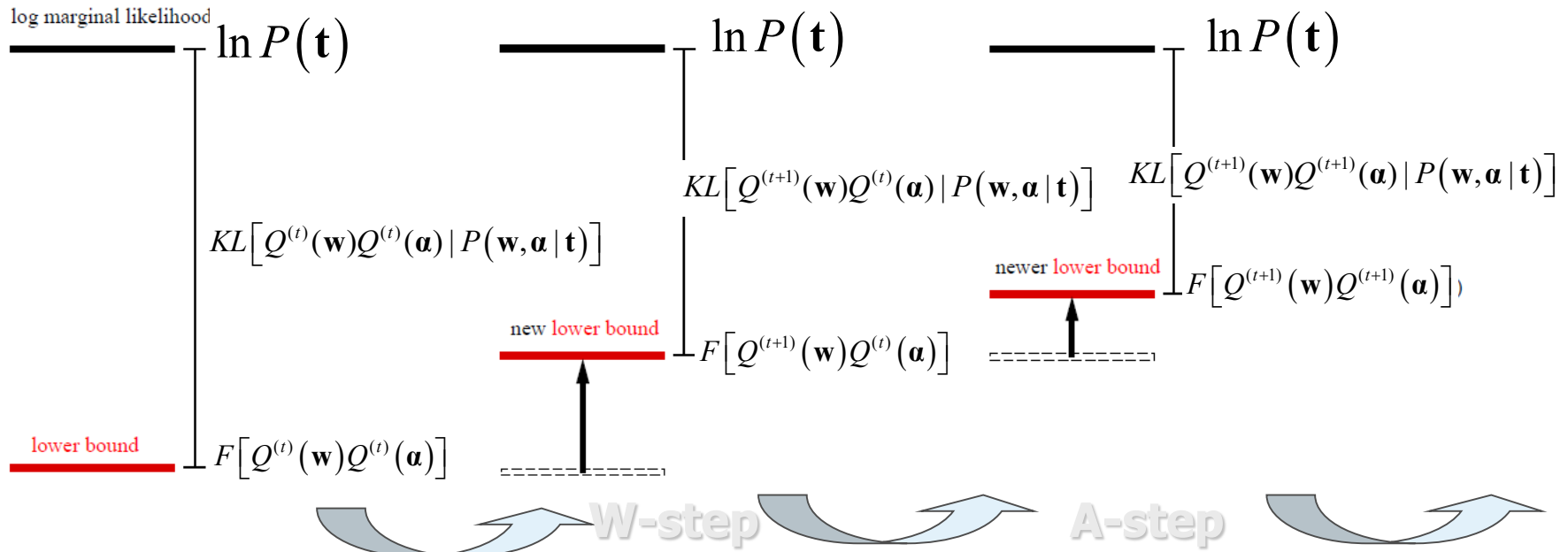
- SLRのパラメータ推定：変分ベイズ法

自由エネルギー $F[Q(\mathbf{w}, \boldsymbol{\alpha})] \equiv \int Q(\mathbf{w}, \boldsymbol{\alpha}) \log(P(\mathbf{t}, \mathbf{w}, \boldsymbol{\alpha}) / Q(\mathbf{w}, \boldsymbol{\alpha})) d\mathbf{w} d\boldsymbol{\alpha}$

$Q(\mathbf{w}, \boldsymbol{\alpha}) = Q(\mathbf{w})Q(\boldsymbol{\alpha})$ の仮定の下、交互に最大化

$F[Q(\mathbf{w}, \boldsymbol{\alpha})] \leq \ln P(\mathbf{t})$ 等号は $Q(\mathbf{w}, \boldsymbol{\alpha}) = P(\mathbf{w}, \boldsymbol{\alpha} | \mathbf{t})$

より、最大値をとる Q は事後分布の良い近似となる。



- SLRのパラメータ推定アルゴリズム

初期化

$$\alpha_i = 1 \quad i = 1, 2, \dots, D$$

W-step

$$[\hat{\mathbf{W}}] = \arg \max_{\mathbf{W}} \langle \log P(\mathbf{t}, \mathbf{W}, \mathbf{a} | \mathbf{X}) \rangle_{\mathbf{a}}$$

$$= \arg \max_{\mathbf{W}} \left[\sum_{i=1}^N \{t_i \log p_i + (1-t_i) \log(1-p_i)\} - \frac{1}{2} \mathbf{W}^t \langle A \rangle_{\mathbf{a}} \mathbf{W} \right]$$

where

$$p_i \square \frac{1}{1 + \exp(-(\mathbf{w}^t \mathbf{x}_i + w_0))}$$

A-step

$$i = 1, 2, \dots, D$$

$$\alpha_i = (1 - \alpha_i S_{ii}) / \hat{w}_i^2$$

$$\text{where } S = (X^t B X + A)^{-1}$$

$$\frac{\partial}{\partial \mathbf{W}} \log P(\mathbf{t} | \mathbf{X}; \mathbf{W}) = \sum_{i=1}^N \{t_i - p_i\} \mathbf{x}_i$$

$$\frac{\partial}{\partial \mathbf{W} \partial \mathbf{W}^t} \log P(\mathbf{t} | \mathbf{X}; \mathbf{W}) = - \sum_{i=1}^N (1-p_i) p_i \mathbf{x}_i \mathbf{x}_i^t$$

$$B = \text{diag}(p_1(1-p_1), \dots, p_N(1-p_N))$$

$$A = \text{diag}(\alpha_0, \dots, \alpha_D)$$

スパースロジスティック判別器

- ロジスティック回帰モデルをベイズ推定に拡張したもの
- スパース事前分布として知られる **Automatic Relevance Determination prior (ARD)** (Mackay 1994, Neal 1996)を適用

ロジスティック回帰モデル

判別モデル

given $(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)$,

$t_i \in \{0, 1\}$: input, \mathbf{x}_i : output

Likelihood : $P(\mathbf{t} | \mathbf{X}; \Theta) = \prod_{i=1}^N p_i^{t_i} (1 - p_i)^{1-t_i}$

where $p_i = 1 / (1 + \exp(-\mathbf{w}^t \mathbf{x}_i))$,

Prior : $P(\theta_k | \alpha_k) \propto N(0, \alpha_k^{-1})$

Hyper Prior: $P(\alpha_k) \propto \alpha_k^{-1}$ (non-informative)

ARD事前分布

推定アルゴリズム

[Tipping, 2001]

W-step

$$[\hat{\mathbf{w}}] = \arg \max_{\mathbf{w}} \left[\sum_{i=1}^N \{t_i \log p_i + (1-t_i) \log(1-p_i)\} - \frac{1}{2} \theta^t \langle A \rangle_{\mathbf{a}} \theta \right]$$

A-step

$i = 1, 2, \dots, D$

$$\alpha_i = (1 - \alpha_i S_{ii}) / \hat{\theta}_i^2$$

where $S = (X^t B X + A)^{-1}$

$$(B)_{ii} = p_i(1-p_i)$$

$$(A)_{ii} = \alpha_i$$

$\langle A \rangle_{\mathbf{a}}$: expectation of A

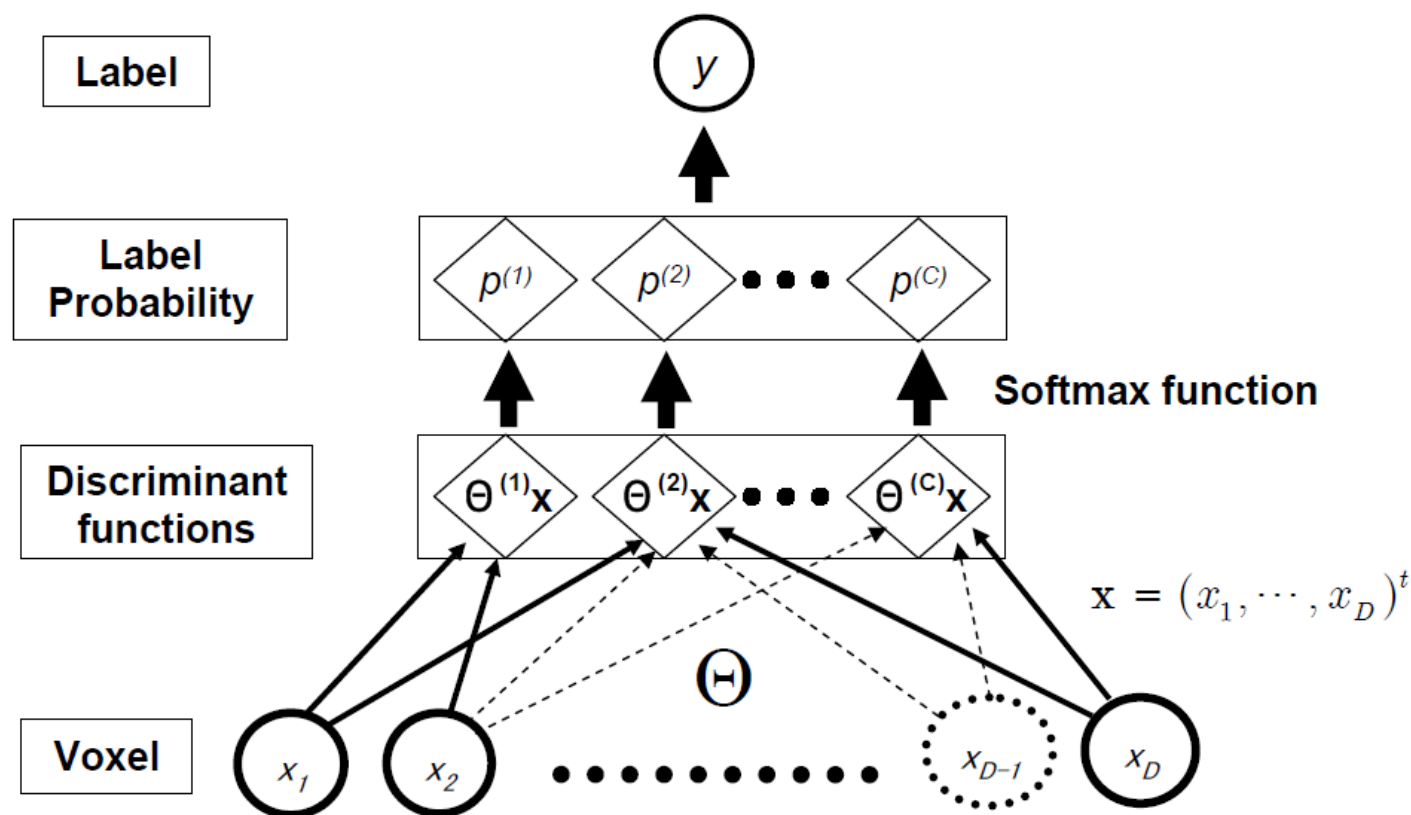
with respect to $\alpha_1, \dots, \alpha_D$

スパースロジスティック判別器

(Yamashita et al. 2008)

多数の特徴量の中から自動的に判別に効果的な特徴量のみを抽出しながら学習する線形判別器

(a)



スパース判別ツールボックス

| | 2値判別 | 多値判別 |
|-----|----------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|
| SLR | SLR-LAP (biclffy_slrlap.m) SLR-VAR (biclffy_slrvar.m) | SMLR (muclffy_smlr.m) SLR-LAP-1vsR (muclffy_slrlapovrm.m) SLR-VAR-1vsR (muclffy_slrvarovrm.m) SLR-VAR-1vs1 (muclffy_slrvarovo.m) ^{New} |

Others classifiers :

- Regularized logistic regression
- Relevant vector machine
- L1-norm-regularized sparse logistic regression

正則化法の確率モデルとしての解釈

$$E = \underbrace{\sum_{n=1}^N \{y_n - (a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_{20} x_n^{20})\}^2}_{\text{データへのフィット}} + \underbrace{\lambda \sum_{d=1}^{20} |a_d|}_{\text{制約条件へのフィット}}$$

データへのフィット

制約条件へのフィット

$$L = \underbrace{\exp\left(-\frac{1}{2} \sum_{n=1}^N \{y_n - (a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_{20} x_n^{20})\}^2\right)}_{\text{尤度関数}} \times \underbrace{\exp\left(-\frac{1}{2} \lambda \sum_{d=1}^{20} |a_d|\right)}_{\text{事前分布}}$$

事後分布

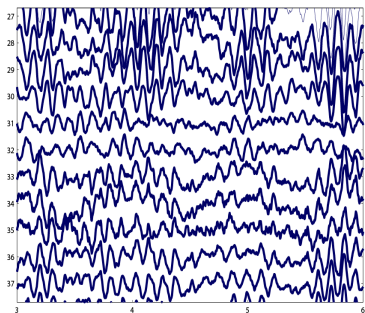
尤度関数

事前分布

ベイズ定理の事後分布最大化として解釈可能

機械学習のプロセス

計測データ

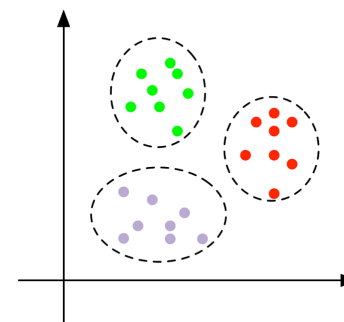
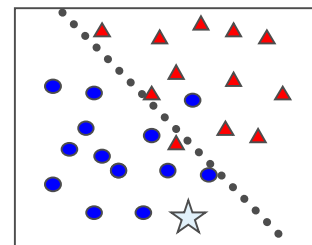


特徴量計算



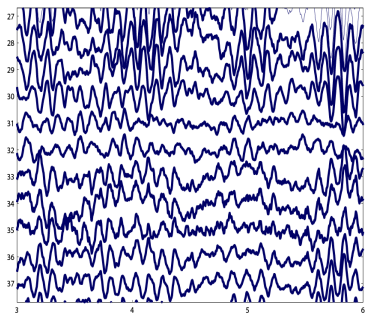
特徴量
ベクトル

判別・回帰
クラスタリング



本講義のターゲット

計測データ

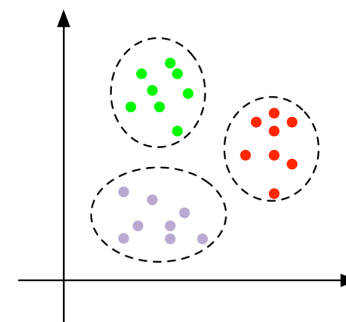
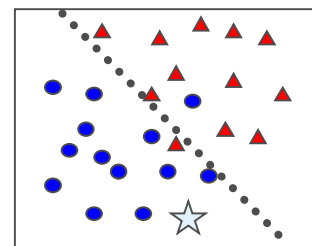


特徴量計算



特徴量
ベクトル

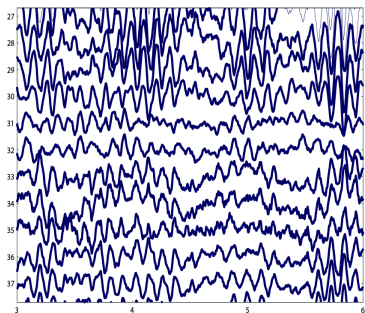
判別・回帰
クラスタリング



モデルの複雑さ
オーバフィット (過適合)
正則化 (事前情報)

深層学習による特徴量の学習 (表現学習)

計測データ

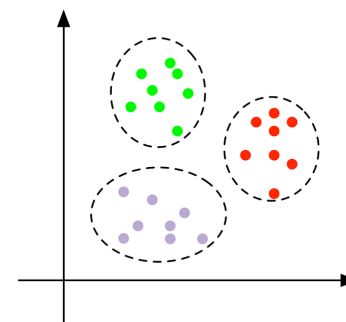
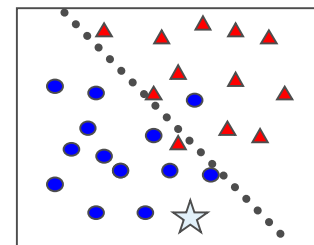


特徴量計算



特徴量
ベクトル

判別・回帰
クラスタリング



深層学習 (表現学習)

ビッグデータにより、計測データからラベルへの関数を複雑なニューラルネットワークを使って学習可能に。