データ解析
第五回「正則化法と判別分析 (手書き文字認識)」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

休講情報

5/20, 6/24 は休講

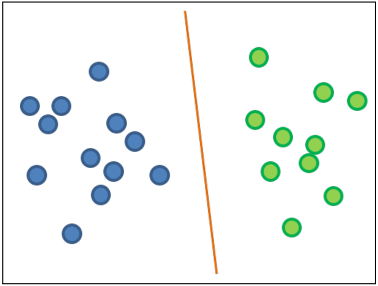
今日の講義内容

- 判別分析
  - LDA
  - ロジスティック回帰
- 正則化法
  - クロスバリデーション
- 手書き文字認識によるデモ

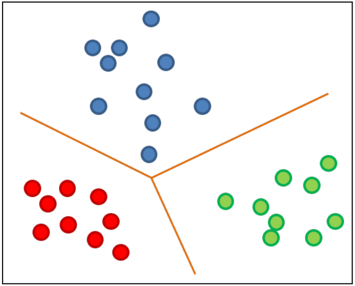
構成

1. 判別分析
2. 正則化法
3. デモ

判別問題



二値判別 (2 ラベル)



多値判別 (3 ラベル以上)

判別分析

データの形式：

$$\big(\underbrace{x_i}_{\text{説明変数}}, \underbrace{y_i}_{\text{ラベル}}\big) \quad (i = 1, \ldots, n).$$

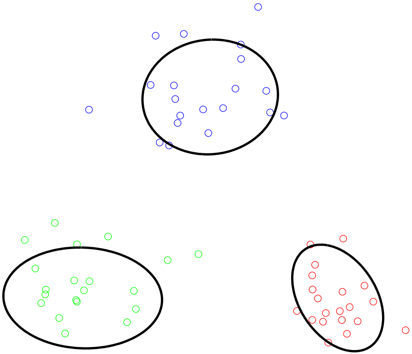
ラベルは 1, 2, ..., K というカテゴリカルな変数.
新しいデータ x がやってきたときに，それがどのラベルに分類されるべきか当てたい.
例：

- 説明変数: 検診データ, ラベル: 疾病のありなし.
- 説明変数: 画像データ, ラベル: 写っている物体.

今日紹介する手法

- LDA (Linear Discriminant Analysis, 線形判別分析)
- ロジスティック回帰

LDA (Linear Discriminant Analysis)



混合ガウス分布

$$P(x) = \sum_{k=1}^{K} \pi_k N(\mu_k, \Sigma_k)$$

いくつかのガウス分布の足しあわせ ($\sum_{k=1}^{K} \pi_k = 1$).
LDA では各カテゴリー (ラベル) を各ガウスコンポーネントに割り当てる.

LDA のモデル

x の周辺分布の密度関数 (y を周辺化):

$$p(x) = \sum_{k=1}^{K} \pi_k g(x|\mu_k, \Sigma_k)$$

(g はガウス分布の密度関数とする)
データは K 個のカテゴリに分けられる．ラベル k は確率 π_k で得られる．

$$Y \sim \mathrm{Mult}(\pi_1, \ldots, \pi_K)$$

カテゴリーが $Y = k$ であるとき，x はガウス分布から得られる:

$$X|\{Y = k\} \sim N(\mu_k, \Sigma_k).$$

では，説明変数 X が与えられたもとでの Y の分布はどうなるだろうか？

ベイズの定理

ベイズの定理

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y = k|X) = \frac{g(X|\mu_k, \Sigma_k)\pi_k}{\sum_{k'=1}^{K} g(X|\mu_{k'}, \Sigma_{k'})\pi_{k'}}.$$

ベイズの定理

ベイズの定理

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y=k|X) = \frac{g(X|\mu_k, \Sigma_k)\pi_k}{\sum_{k'=1}^{K} g(X|\mu_{k'}, \Sigma_{k'})\pi_{k'}}.$$

LDA では $\underline{\Sigma_k \text{ がすべての } k \text{ で等しい}}$ と仮定し，μ_k, Σ_k, π_k を最尤推定:

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i, \quad \hat{\Sigma}(=\hat{\Sigma}_k) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{y_i})(x_i - \hat{\mu}_{y_i})^\top, \quad \hat{\pi}_k = \frac{n_k}{n}.$$

新しいデータ X に対しては次の式で分類:

$$\hat{Y} = \mathop{\arg\max}_{k=1,\dots,K} \frac{g(X|\hat{\mu}_k, \hat{\Sigma})\hat{\pi}_k}{\sum_{k'=1}^{K} g(X|\hat{\mu}_{k'}, \hat{\Sigma})\hat{\pi}_{k'}}.$$

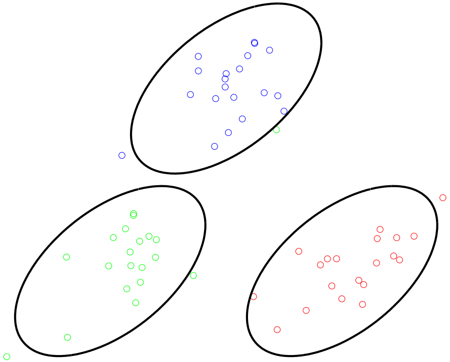
寄与率の高いカテゴリに分類されやすい．

マハラノビス距離

実は，$\hat{Y}$ は マハラノビス距離最小化 で求まる:

$$\hat{Y} = \arg\min_k \ (X - \mu_k)^\top \hat{\Sigma}^{-1}(X - \mu_k) - 2\log(\hat{\pi}_k).$$

これは分散一定 ($\Sigma_k = \Sigma$ $(\forall k)$) の仮定による (チェックせよ).
判別平面は線形になる.



QDA (quadratic discriminant analysis)

Σ_k を k に依存して決めるモデル.

LDA と違うのは

$$\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top,$$

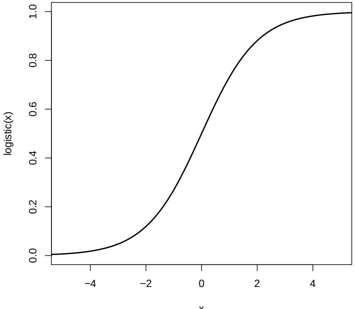
とする部分だけ.
ただし，判別平面が線形ではなくなる.

ロジスティック回帰

前回の一般化線形モデルを参照.
二値判別モデル:

$$P(Y = 1|x) = \frac{1}{1 + \exp(-\beta^\top x)},$$
$$P(Y = 2|x) = \frac{1}{1 + \exp(\beta^\top x)}.$$

（前回は $Y = 0, 1$ と書いていたが，今回は $Y = 1, 2$ で書く）



ロジスティック回帰

前回の一般化線形モデルを参照.
二値判別モデル:

$$P(Y=1|x) = \frac{1}{1+\exp(-\beta^\top x)},$$
$$P(Y=2|x) = \frac{1}{1+\exp(\beta^\top x)}.$$

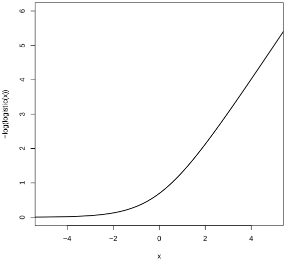
（前回は $Y=0,1$ と書いていたが，今回は $Y=1,2$ で書く）

二値判別の対数尤度最大化:

$$\ell(Y,\beta^\top x) := \begin{cases} \log(1+\exp(-\beta^\top x)), & (Y=1), \\ \log(1+\exp(\beta^\top x)), & (Y=2). \end{cases}$$

としたとき，

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i).$$



ロジスティック回帰: 多値判別モデル

多値判別モデル:

$$P(Y=k|x) = \frac{\exp(\beta_k^\top x)}{1+\sum_{k'=1}^{K-1}\exp(\beta_{k'}^\top x)} \quad (k<K),$$
$$P(Y=K|x) = \frac{1}{1+\sum_{k'=1}^{K-1}\exp(\beta_{k'}^\top x)}.$$

※ $K=2$ の時はさきほどの二値判別モデルと同値になることを確かめよ.

負の対数尤度最小化: $\beta = [\beta_1,\dots,\beta_{K-1}] \in \mathbb{R}^{d\times(K-1)}$ に対して,

$$\ell(Y,\beta^\top X) = -\log(P(Y|X))$$
$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell(y_i,\beta^\top x_i).$$

これも凸最適化で解ける.

構成

1 判別分析

2 正則化法

3 デモ

正則化法

普通のロス関数 (負の対数尤度) 最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i).$$

正則化付きロス関数最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \underbrace{\lambda \|\beta\|^2}_{\text{正則化項}}.$$

※ 正則化項は二乗ノルム以外にもいろいろある (例: ℓ_1-ノルム)

正則化法

普通のロス関数 (負の対数尤度) 最小化:

$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i).$$

正則化付きロス関数最小化:

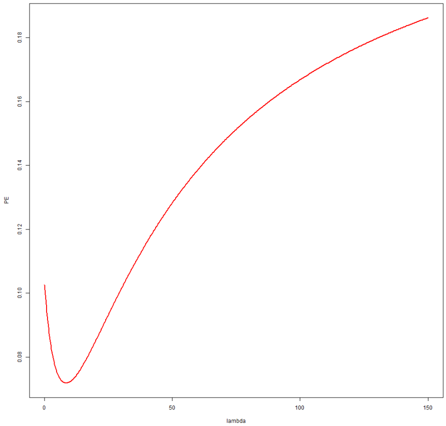
$$\min_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \underbrace{\lambda \|\beta\|^2}_{\text{正則化項}}.$$

※ 正則化項は二乗ノルム以外にもいろいろある (例: ℓ_1-ノルム)

- 正則化項をつけることで分散が抑えられ，特に高次元データ解析で安定した精度が得られる．
- その分，バイアスが乗る．

→ 適切な正則化の強さ (λ) を選ぶ必要がある．

$n = 100,\ d = 10$ のリッジ回帰 (ガウスマルコフモデル+二乗ノルム正則化)



正則化定数 (λ) vs 予測誤差 ($\mathrm{E}_X[({\beta^*}^\top X - \hat{\beta}^\top X)^2]$)

$$\ell(y, \beta^\top x) = \frac{1}{2\sigma^2}(y - \beta^\top x)^2,$$

$$\sum_{i=1}^{n} \ell(y_i, \beta^\top x_i) + \lambda \|\beta\|^2 = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^2 + \lambda \|\beta\|^2.$$

クロスバリデーション

クロスバリデーション **(CV, cross validation)**: 適切な正則化定数を選ぶ方法.

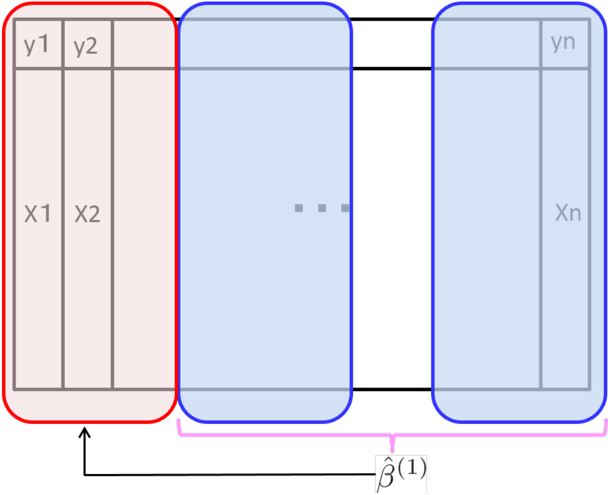
観測データへの当てはまりではなく予測誤差を最小化.
観測データへの当てはまりを最良にするのは $\lambda = 0$.

k-fold クロスバリデーション

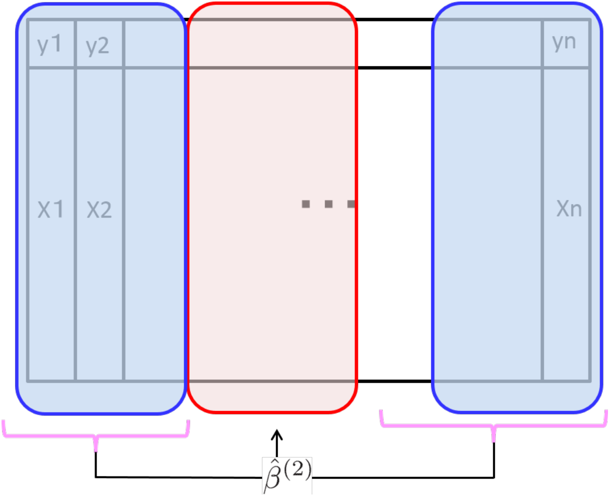
1. まずデータを k 個に分割する.
2. 分割したデータの一つをテストデータとしてとっておき，残りのデータで推定.
3. テストデータ上での予測誤差を計算.
4. 手順 2-3 を k 個のテストデータの取り方について繰り返す.
5. k 回繰り返しの予測誤差の平均を取る=CV スコア.

CV スコアを最小にする λ を選べば良い.

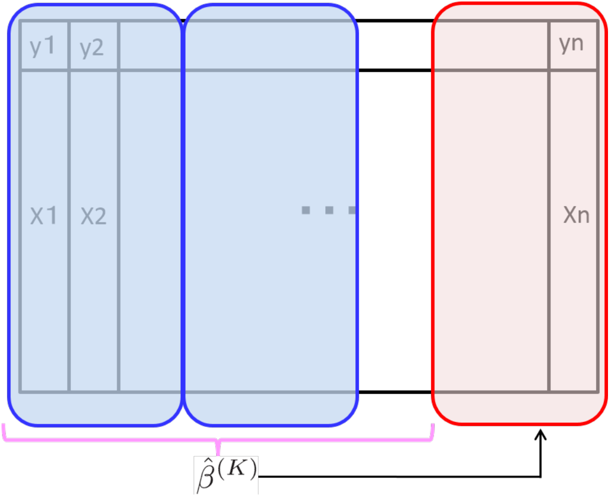
特に $k = n$ (サンプル数) の時，Leave-One-Out-CV (LOOCV) と呼ぶ.



$$\frac{1}{|I_1|}\sum_{i\in I_1}\ell(y_i,\hat{\beta}^{(1)\top}x_i)$$



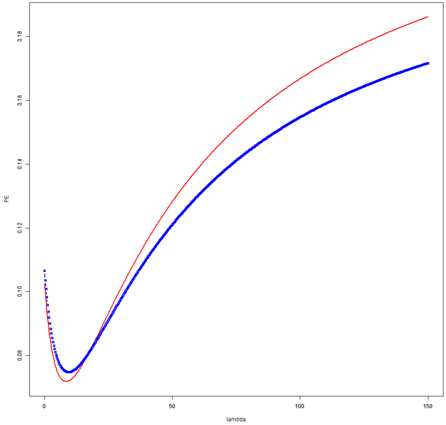
$$\frac{1}{|I_2|}\sum_{i\in I_2}\ell(y_i, \hat{\beta}^{(2)\top} x_i)$$



$$\frac{1}{|I_K|}\sum_{i\in I_K}\ell(y_i,\hat{\beta}^{(K)\top}x_i)$$

実例

$n = 100,\ d = 10$ のリッジ回帰 (ガウスマルコフモデル+二乗ノルム正則化)



予測誤差 (赤線) と CV スコア (青線)

構成

1 判別分析

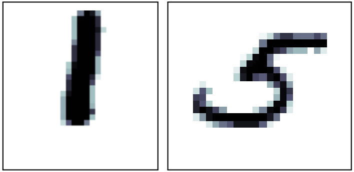
2 正則化法

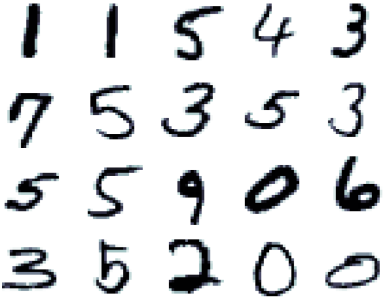
3 デモ

手書き文字認識

MNIST 手書き文字データ:

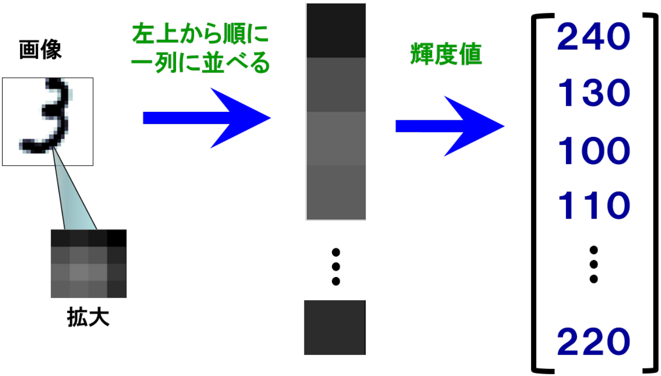
- 28 × 28 のグレースケール画像.
- 6000 個の訓練サンプル, 10000 個のテストサンプル.





※ 講義情報ページから csv ファイルを入手可能.

データ形式



輝度値は 0 から 255 の整数値.

講義情報ページ

`http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html`