

データ解析

第六回 「L1 正則化法：高次元データ解析」

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

7/7 は休講

今日の講義内容

- L_1 正則化
- バイオデータとスパムメール判別によるデモ
- レポート課題

モデル選択の難しさ

高次元データ：パラメータの次元 p が大きい.

AIC でモデル選択

→ 2^p 通り！ (NP 困難)

→ 計算に時間がかかりすぎる

→ L_1 正則化：凸最適化で変数選択

(R の `step()` では変数を一つ加えたり削ったりして最適なモデルを探索. 初期値に大きく依存する.)

スパース推定 [Lasso]

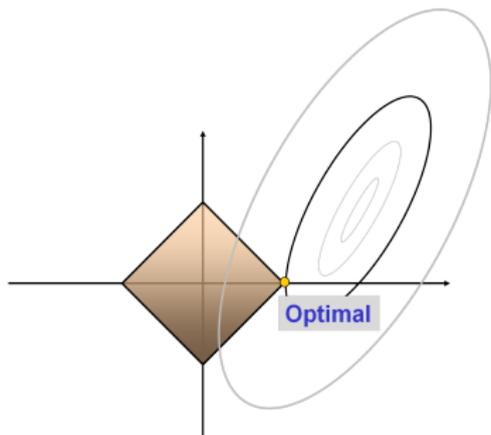
デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$.

p (次元) $\gg n$ (サンプル数).

真のベクトル $\beta^* \in \mathbb{R}^p$: 非ゼロ要素の個数がたかだか d 個 (スパース).

モデル: $Y = X\beta^* + \xi$.

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^p |\beta_j|.$$



スパース推定 [Lasso]

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$.

p (次元) $\gg n$ (サンプル数).

真のベクトル $\beta^* \in \mathbb{R}^p$: 非ゼロ要素の個数がたかだか d 個 (スパース).

モデル: $Y = X\beta^* + \xi$.

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^p |\beta_j|.$$

Theorem (Lasso の収束レート (Bickel et al., 2009, Zhang, 2009))

デザイン行列が *Restricted eigenvalue condition* Bickel et al. (2009) かつ

$\max_{i,j} |X_{ij}| \leq 1$ を満たし, ノイズが $E[e^{\tau \xi_i}] \leq e^{\sigma^2 \tau^2 / 2}$ ($\forall \tau > 0$) を満たすなら, 確率 $1 - \delta$ で

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \frac{d \log(p/\delta)}{n}.$$

※次元が高くて, たかだか $\log(p)$ でしか効いてこない. 実質的な次元 d が支配的.

一般化加法モデルへの L_1 正則化の適用

L_1 正則化 :

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i^\top \beta) + \lambda_n \sum_{j=1}^p |\beta_j|.$$

L_2 正則化 (前回の授業参照. Ridge regression はこれに含まれる) :

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i^\top \beta) + \lambda_n \sum_{j=1}^p \beta_j^2.$$

glmnet で L_1 正則化

glmnet(x, y, ..., alpha = 1)

alpha (デフォルト 1) で正則化項を L_1 と L_2 の間で調整 :

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i^\top \beta) + \lambda_n \sum_{j=1}^p (\alpha |\beta_j| + \frac{1}{2} (1 - \alpha) \beta_j^2).$$

この α の値を alpha で指定. (これを Elasticnet 正則化と呼ぶ)

$\alpha = 1 \rightarrow L_1$ 正則化

$\alpha = 0 \rightarrow L_2$ 正則化

LiblineR で L_1 正則化

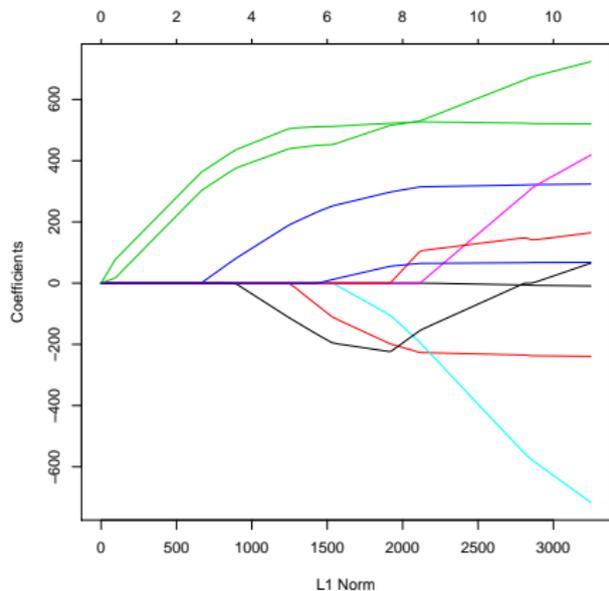
LiblineR(x,y, type=6)

type で正則化の種類とロス関数を指定。デフォルトは type=0.

- 0 - L2-regularized logistic regression
- 1 - L2-regularized L2-loss support vector classification (dual)
- 2 - L2-regularized L2-loss support vector classification (primal)
- 3 - L2-regularized L1-loss support vector classification (dual)
- 4 - multi-class support vector classification by Crammer and Singer
- 5 - L1-regularized L2-loss support vector classification
- 6 - L1-regularized logistic regression
- 7 - L2-regularized logistic regression (dual)

※ LiblineR は判別問題しか扱えない。

```
library(glmnet)
library(lars)
data(diabetes) #糖尿病データ. 線形回帰問題. lars パッケージに入っている.
lasso <- glmnet(diabetes$x, diabetes$y)
plot(lasso)
```



左図の意味：

λ を動かしていった時に、推定された係数がどうふるまうか。

λ が大きい時、つまり L_1 ノルムが小さい時は推定量がスパースになっていることがわかる。

正則化パス。横軸は推定量の L_1 ノルム。

スパムメール判別

メールをスパムかそうでないかを判別したい。

判別分析の枠組みに乗せるため、各メールを一つのベクトル x として表したい。

Bag of words: 出現した単語の頻度を並べたベクトル。

各要素が各単語の頻度に対応。

単語数分だけの次元になるため高次元になりやすい。自然言語処理では 100 万次元はザラ。

$$\begin{pmatrix} \text{"please" の出現頻度} \\ \text{"credit" の出現頻度} \\ \vdots \\ \text{"money" の出現頻度} \end{pmatrix}$$

今回は UCI Machine Learning Repository の Spam e-mail database を利用。
これは 57 次元と次元は低い方。

レポート課題三回目

- ① Ridge 回帰で k-fold CV を実行する関数を書け。入力は $Y \in \mathbb{R}^n$ (従属変数), $X \in \mathbb{R}^{n \times d}$ (説明変数), λ (正則化定数), k (k-fold CV) で, 出力は二乗誤差 $((y - x^T \hat{\beta})^2)$ の CV スコア。ただし, ライブラリを用いるのではなく, 推定量の計算から CV スコアの計算まで自分でコーディングすること。
- ② 講義第五回に用いた手書き文字認識データで, 正則化パラメータと判別精度の関係をグラフにせよ。その際, サンプル数をいくつか変えてみて, それらのグラフを重ね書きせよ。最良な判別精度を達成する正則化パラメータはサンプル数に依存してどう変化しているか?
- ③ 上のデータにさらに CV スコアを重ね書きせよ。LiblineR では `cross=k` とおくことで k-fold CV スコアが得られる。
`cvscore <- Liblinear(..., cross = 10)`
- ④ 自分が興味のあるデータにおいて, 最小二乗法以外のこれまで講義で紹介した手法を用いて解析せよ。線形判別, 正則化付き線形判別, 一般化線形モデル, 正則化付き線形回帰など, 最小二乗法以外の説明変数から従属変数を予測する問題ならなんでも良い。なお, 講義で使わなかった関数を用いても構わない。また, 最小二乗法以外の手法を用いても, 結局最小二乗法が一番 (何らかの意味で) 良かった場合はその旨論じてもよい。自分で工夫していればその分評価は上がる。

レポートの提出方法

- 私宛にメールにて提出.
- 件名に 必ず 「データ解析第 n 回レポート」と明記し, R のソースコードと結果をまとめたレポートを送付のこと.
- 氏名と学籍番号も忘れず明記すること.
- レポートは本文に載せても良いが, pdf などの電子ファイルにレポートを出力して添付ファイルとして送付することが望ましい (これを期に tex の使い方を覚えることを推奨します).
- 提出期限は講義最終回まで.

※相談はしても良いですが, コピペは厳禁です.

講義情報ページ

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

- P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.
- T. Zhang. Some sharp performance bounds for least squares regression with l_1 regularization. *The Annals of Statistics*, 37(5):2109–2144, 2009.