The Elements of Statistical Learning
Chap.18: High-Dimensional Problems: $p \gg N$

Kosuke Kito

August 28, 2020

Section 1

Introduction
p.653-

本日のお題 - $p \gg N$ 問題

特徴量の数がサンプル数よりもずっと大きいとき ($p \gg N$) に困っちゃう話.

- 困っちゃうポイントは, high variance と overfitting
- simple, highly regularized な手法が使われる.
- 主な話題は以下の２つ.
  - prediction
  - feature selection, assesment

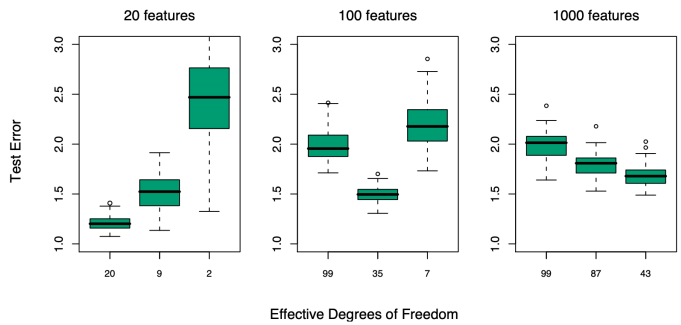


FIGURE 18.1. *Test-error results for simulation experiments. Shown are box-plots of the relative test errors over* 100 *simulations, for three different values of* p*, the number of features. The relative error is the test error divided by the Bayes error,* σ^2*. From left to right, results are shown for ridge regression with three different values of the regularization parameter* λ*:* 0.001, 100 *and* 1000*. The (average) effective degrees of freedom in the fit is indicated below each plot.*

流れ

- 流れを書く.

Section 2

LDA の正則化 - Diagonal LDA と NSC
p.651-

LDA の復習 1 - コンセプト

$p \gg N$ 問題の最初の回避策は, “Diagonal LDA” という線型判別法の強烈な正則化バージョン.
とりあえず, LDA の復習. (不要であれば飛ばします.)

- 分類のための手法.
- 各入力 x に対して, 事後確率 $\Pr[k \mid X = x]$ が最大になるクラス k をクラスの推定値とする.
- 各クラス内で, 入力変数は多変量ガウス分布に従うと仮定.
- 各クラスのクラス内分散が等しいと仮定.
  →クラス間の境界が線形になる.
- 数式で書くと次ページの流れ.

LDA の復習 2 - 定式化と計算

- 各クラス内の分布はガウス分布と仮定.

$$\Pr[X = x|G = k] = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^{\mathrm{T}}\Sigma(x-\mu_k)}$$

- ベイズの定理より事後確率は以下.

$$\Pr[k \mid X = x] = \frac{\Pr[X = x|G = k]\Pr[G = k]}{\sum_l \Pr[X = x|G = l]\Pr[G = l]}$$

- 事後確率の大小比較のため対数比 (log-ratio) を見る.

$$\begin{aligned}
\log \frac{\Pr[k \mid X = x]}{\Pr[l \mid X = x]} &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^{\mathrm{T}}\Sigma^{-1}(\mu_k - \mu_l) + x^{\mathrm{T}}\Sigma^{-1}(\mu_k - \mu_l) \\
&= (\log \pi_k - \frac{1}{2}\mu_k^{\mathrm{T}}\Sigma^{-1}\mu_k + x^{\mathrm{T}}\Sigma^{-1}\mu_k) \\
&\quad -(\log \pi_l - \frac{1}{2}\mu_l^{\mathrm{T}}\Sigma^{-1}\mu_l + x^{\mathrm{T}}\Sigma^{-1}\mu_l)
\end{aligned}$$

- 結局, 点 x がクラス k である度合い (discriminant score) は以下を評価すれば良い.

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

Diagonal LDA - 線型判別法の強烈な正則化バージョン

- 基本的なコンセプトは, 前述の LDA と同じ. 以下の条件を追加する.

$$\Sigma = \mathrm{diag}(s_1, s_2, \ldots, s_p) \quad \text{(対角行列)}$$

- すると, discriminant score は, (クラスに依らない定数 $-x^{\mathrm{T}}\Sigma^{-1}x$ を足して 2 倍することで,) 以下になる.

$$\delta_k(x) = -\sum_{j=1}^{p} \frac{(x_j - \bar{x}_{kj})^2}{s_j^2} + 2\log \pi_k$$

- この discriminant score を使って, 以下のルールで分類する.

$$C(x) = \arg\max_l \delta_l(x)$$

Diagonal LDA - 線型判別法の強烈な正則化バージョン

Diagonal LDA について何点か補足.

- discriminant score は距離に見える.
  → Diagonal LDA は, 適当な標準化をしたデータにおける nearest centroid 法のようにも見える.
- 変数間の共分散が 0 という仮定を, 独立律 (independent rule) ともいう.
- 高次元の時には, effective なことが多いらしい.
- この方法の欠点の一つは, 特徴量選択 (feature selection) ができないこと. 高次元の入力の時には, 一部の変数を選び出せる方法を使いたい.
  →もっと正則化の条件を強めるとパフォーマンスがさらに上がるらしい.
- 次は, 特徴量選択が行われるような正則化の条件をかけたバージョンを考えます.

Nearest Shrunken Centroids

前出の Diagonal LDA = Nearest Centroid 法の centroid を縮小 (shrinkage) させることで, 特徴量選択を行えるようにする.

- 基本的な計算は, Diagonal LDA と一緒.
- discriminant score の計算に使う centroid を単純な平均 $\bar{x}_{kj}$ から変える.
- まず, あるパラメータ X_j のクラス k 内での平均 $\bar{x}_{kj}$ と全体での平均 $\bar{x}_j$ の差を標準化する.

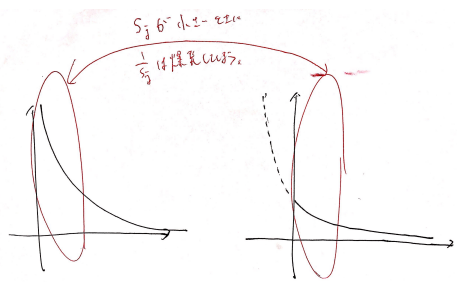
$$d_{kj} = \frac{\bar{x}_{kj} - \bar{x}_j}{m_k(s_j + s_0 b)}$$

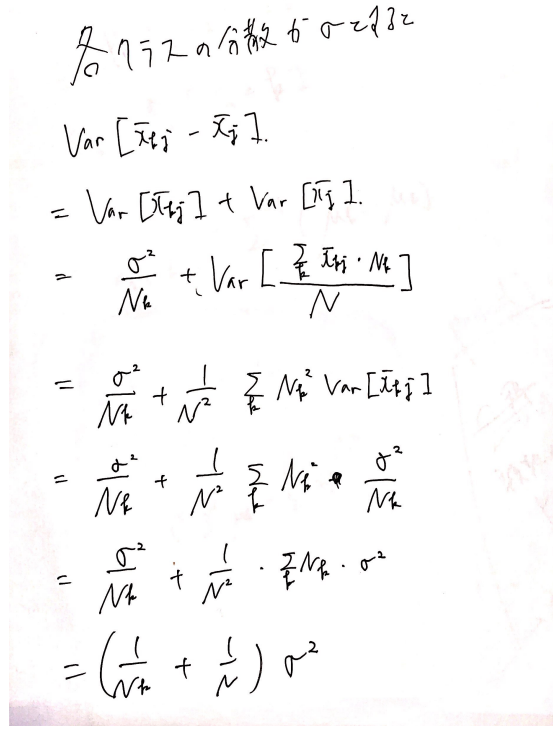
ただし, 各項は以下.

$m_k = \frac{1}{N_k} - \frac{1}{N}$: 疑問点

s_0 = 小さな定数

s_jが小さい時に d_{kj}が大きくなりすぎないように





Nearest Shrunken Centroids

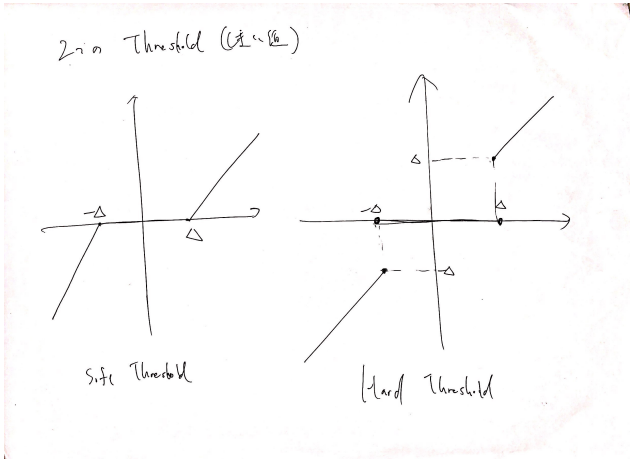
- この標準化された距離を soft-threshold

$$d'_{kj} = \mathrm{sign}(d_{kj})(|d_{kj}| - \Delta)$$

もしくは hard-threshold

$$d'_{kj} = d_{kj} I(|d_{kj}| \geq \Delta)$$

で縮小させる.



- すなわち, 以下.

$$\bar{x}'_{kj} = \bar{x}_j + m_k(s_j + s_0)d'_{jk}$$

寄与の小さい特徴量を無視するようになっている.

- Diagonal LDA の discriminant score の $\bar{x}_{kj}$ の代わりに, $\bar{x}'_{kj}$ を使えば, NSC の完成.

Section 3

2 次で正則化した線型分類

p.654-

Regularized Discriminant Analysis

- 判別分析の正則化を考える.
- 以前見た正則化は, LDA と QDA でバランスを取るために, 分散行列を以下で計算した.

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

- 今回は違うバージョン. 対角行列に近づけようとする.

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\mathrm{diag}(\hat{\Sigma})$$

- 上記の正則化で, $\gamma = 0$ のときは, Diagonal LDA, すなわち, 縮小のない NSC と同じ.
- ridge 回帰が分散共分散行列を対角行列に近づけようとするのと, 似ている.

$$\hat{\beta}^{\mathrm{ridge}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- 線型判別を, 各クラスに数値を割り当てた線型回帰と思うと, ridge 回帰との関連をより正確に記述できるそう.

ロジスティック回帰

- 以下の式を使ってきた.

$$\Pr[G = k \mid X = x] = \frac{\exp(\beta_{k0} + x^{\mathrm{T}}\beta_k)}{\sum_l \exp(\beta_{l0} + x^{\mathrm{T}}\beta_l)}$$

- 対数尤度に正則化項を加えた以下を最大化する.

$$\sum_{i=1}^{N} \log \Pr[g_i \mid x_i] - \frac{\lambda}{2} \sum_{k=1}^{K} ||\beta_k||^2$$

- 1 個目の定式化だと, over-parametrized だけど, 正則化項のおかげでいい感じ.
- 定数項のみいい感じじゃないので, 適当に条件加えよう.

$$\frac{e^3}{e^1+e^2+e^3} = \frac{e^3 \times e^{-2}}{(e^1+e^2+e^3)\times e^{-2}}$$

$$= \frac{e^1}{e^{-1}+e^0+e^1}$$

Logistic 回帰の redundancy

ロジスティック回帰

- 前述の最大化問題は凸なので, Newton 法とかで解ける.
- separable なデータに対して, $\lambda \to 0$ とすると, マージン最大化と同じ結果になる.(?)
  → SVM もうまく関連付けられそう.

Support Vector Classifier

- 前に出てきた話。マージン最大化。
- $p \gg N$ のとき、ほぼ確で線型分離可能なので、attractive.
- 正則化しなくても有効。頑張って正則化しても、正則化なしと同程度のパフォーマンスのことが多い。
- 多数のクラスへの分類への応用方法を２つ紹介。
  - one versus one (ovo)
    全ての２つのクラスの組み合わせ ($K(K-1)/2$ 通り) 全てについて、SVM で分類する。
    点 x について、上記の分類全ての結果、最も多く分類されるクラスを推定値とする。
  - one versus all (ova)
    各クラスとそのクラス以外に分けて SVM で分類する。
    教会からの符号付き距離 (confidence) が最も大きいクラスを推定値とする。
- 正則化ロジスティック回帰と近しい結果を返す。

特徴量選択

- p が大きい時、特徴量選択は重要。解釈可能性のため。
- DLDA, LR, SVC は、アルゴリズム内に特徴量選択の機能を含まない。(二次正則化のため)
  →外付けの特徴量選択手法が提案されている。
- 例。Recursive Feature Elimination.
  重みの小さい特徴量から無視していく。
  →あまり上手くいかないらしい。
  (“ we do not have an explanation for this behavior.”)
- Kernel 法使って外れ値に強くさせることも可能。

計算の工夫

$p \gg N$ で二次正則化を考えた時に使える計算の工夫について。

- a