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

Kosuke Kito

August 27, 2020

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

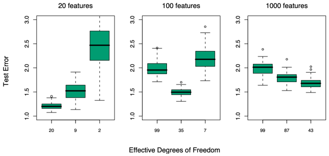

FIGURE 18.1. *Test-error results for simulation experiments. Shown are boxplots of the relative test errors over 100 simulations, for three different values of p, the number of features. The relative error is the test error divided by the Bayes error, σ^2. From left to right, results are shown for ridge 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 1

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規化バージョン．
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規化バージョン

- 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規化バージョン

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。
  →もっと正規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規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。

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

$s_0 =$ 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

各クラスの分散がσとすると

$$\begin{aligned}
&\mathrm{Var}[\bar{x}_{kj} - \bar{x}_j] \\
&= \mathrm{Var}[\bar{x}_{kj}] + \mathrm{Var}[\bar{x}_j] \\
&= \frac{\sigma^2}{N_k} + \mathrm{Var}\left[\frac{\sum_k \bar{x}_{kj} \cdot N_k}{N}\right] \\
&= \frac{\sigma^2}{N_k} + \frac{1}{N^2} \sum_k N_k^2 \mathrm{Var}[\bar{x}_{kj}] \\
&= \frac{\sigma^2}{N_k} + \frac{1}{N^2} \sum_k N_k^2 \cdot \frac{\sigma^2}{N_k} \\
&= \frac{\sigma^2}{N_k} + \frac{1}{N^2} \cdot \sum_k N_k \cdot \sigma^2 \\
&= \left(\frac{1}{N_k} + \frac{1}{N}\right) \sigma^2
\end{aligned}$$