先端データ解析論　第 5 回小レポート

情報理工学系研究科電子情報学専攻 M1 堀 紡希 48216444

2021 年 5 月 11 日

宿題 1

線形モデルでの最小二乗分類における計画行列が

$$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$$

であるので, 最小二乗分類による識別境界は

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

である. 一方

$$\mathbf{0} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i = \frac{1}{n_+ + n_-} \left(\sum_{i:y_i=+1} x_i + \sum_{i:y_i=-1} x_i \right) = \frac{1}{n_+ + n_-} (n_+ \hat{\boldsymbol{\mu}}_+ + n_- \hat{\boldsymbol{\mu}}_-)$$

したがって $n_+ \hat{\boldsymbol{\mu}}_+ + n_- \hat{\boldsymbol{\mu}}_- = 0$ が成り立つ.
また以下の関係式が成り立つ.

$$\begin{aligned} \boldsymbol{X}^\top \boldsymbol{y} &= (y_1 \boldsymbol{x}_1 + \cdots + y_n \boldsymbol{x}_n) \\ &= n_+ \hat{\boldsymbol{\mu}}_+ + n_- \hat{\boldsymbol{\mu}}_- = 2 n_+ \hat{\boldsymbol{\mu}}_+ \end{aligned}$$

さらに

$$\boldsymbol{X}^\top \boldsymbol{X} = (\boldsymbol{x}_1 \boldsymbol{x}_1^\top + \cdots + \boldsymbol{x}_n \boldsymbol{x}_n^\top)$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\mu})(\boldsymbol{x}_i - \hat{\mu})^\top = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

から,

$$\boldsymbol{X}^\top \boldsymbol{X} = (n_+ + n_-) \hat{\boldsymbol{\Sigma}}$$

が成り立つ.

以上から, 最小二乗分類の境界線は以下で表せる.

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{2n_+}{n_+ + n_-} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_+$$

これはフィッシャー判別分析による識別境界,

$$\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_+ + \hat{\boldsymbol{\mu}}_-) = 2\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}_+$$

のスカラー倍 ($= n_+/(n_+ + n_-)$ 倍) になっている.

宿題 2

別の ipynb ファイルで提出します.