判別分析 - 評価

数理科学続論J

(Press `?` for help, `n` and `p` for next and previous slide)

村田 昇

2019.12.06

講義の予定

- 第1日: 判別分析の考え方
- **第2日: 判別分析の評価**

判別分析の復習

3.1

判別分析

- 個体の特徴量から その個体の属するクラスを予測する関係式を構成
- **事前確率** (prior probability): $\pi_k = P(Y = k)$
  - $X = \boldsymbol{x}$ が与えられる前に予測されるクラス
- **事後確率** (posterior probability): $p_k(\boldsymbol{x})$
  - $X = \boldsymbol{x}$ が与えられた後に予測されるクラス

$$p_k(\boldsymbol{x}) := P(Y = k | X = \boldsymbol{x})$$

  - 所属する確率が最も高いクラスに個体を分類

判別関数

- 判別の手続き
  - 特徴量 $X = \boldsymbol{x}$ の取得
  - 事後確率 $p_k(\boldsymbol{x})$ の計算
  - 事後確率最大のクラスにデータを分類
- **判別関数**: $\delta_k(\boldsymbol{x})$ $(k = 1, \dots, K)$

$$p_k(\boldsymbol{x}) < p_l(\boldsymbol{x}) \Leftrightarrow \delta_k(\boldsymbol{x}) < \delta_l(\boldsymbol{x})$$

  事後確率の順序を保存する計算しやすい関数
- 判別関数 $\delta_k(\boldsymbol{x})$ を最大化するようなクラス k に分類

線形判別

- $f_k(\boldsymbol{x})$ の仮定:
  - q 変量正規分布の密度関数
  - 平均ベクトル $\boldsymbol{\mu}_k$: クラスごとに異なる
  - 共分散行列 Σ: すべてのクラスで共通

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{q/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k\right.$$

- 線形判別関数: $\boldsymbol{x}$ の1次式

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \Sigma^{-1} \boldsymbol{\mu}_k + \log \pi_k$$

2次判別

- $f_k(\boldsymbol{x})$ の仮定:
  - q 変量正規分布の密度関数
  - 平均ベクトル $\boldsymbol{\mu}_k$: クラスごとに異なる
  - 共分散行列 Σ_k: **クラスごとに異なる**

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{q/2}\sqrt{\det \Sigma_k}} \exp\biggl(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}$$

- 2次判別関数: $\boldsymbol{x}$ の2次式

$$\delta_k(\boldsymbol{x}) = -\frac{1}{2} \det \Sigma_k - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) + \log$$

Fisherの線形判別

- 新しい特徴量 $Z = \boldsymbol{\alpha}^\top X$ を考える
- 良い Z の基準:
  - クラス内では集まっているほど良い
  - クラス間では離れているほど良い
- Fisherの基準:

$$\text{maximize} \quad \boldsymbol{\alpha}^\top B \boldsymbol{\alpha} \quad \text{s.t.} \quad \boldsymbol{\alpha}^\top W \boldsymbol{\alpha} = \text{const.}$$

- $\boldsymbol{\alpha}$ は $W^{-1}B$ の第1から第 $K-1$ 固有ベクトル
- 判別方法: 特徴量の距離を用いる
  - $d_k = \sum_{l=1}^{K-1} (\alpha_l^\top \boldsymbol{x} - \alpha_l^\top \mu_k)^2$ が最小のとなるクラス k に判別

2値判別分析の評価

4.1

誤り率

- 単純な誤り:

$$(\text{誤り率}) = \frac{(\text{誤って判別されたデータ数})}{(\text{全データ数})}$$

- 判別したいラベル: 陽性 (positive)
  - 正しく陽性と判定: **真陽性** (true positive; TP)
  - 誤って陽性と判定: **偽陽性** (false positive; FP) (第I種過誤)
  - 誤って陰性と判定: **偽陰性** (false negative; FN) (第II種過誤)
  - 正しく陰性と判定: **真陰性** (true negative; TN)

	真値は陽性	真値は陰性

混同行列 (confusion matrix)

	真値は陽性	真値は陰性
判別は陽性	真陽性 (True Positive)	偽陽性 (False Positive)
判別は陰性	偽陰性 (False Negative)	真陰性 (True Negative)

(転置で書く流儀もあるので注意)

	判別は陽性	判別は陰性

混同行列

	判別は陽性	判別は陰性
真値は陽性	真陽性 (True Positive)	偽陰性 (False Negative)
真値は陰性	偽陽性 (False Positive)	真陰性 (True Negative)

(パターン認識や機械学習で多く見られた書き方．誤差行列 (error matrix) ともいう)

いろいろな評価基準

$$(\text{真陽性率}) = \frac{TP}{TP + FN} \qquad \text{(true positive rate)}$$

$$(\text{真陰性率}) = \frac{TN}{FP + TN} \qquad \text{(true negative rate)}$$

$$(\text{適合率}) = \frac{TP}{TP + FP} \qquad \text{(precision)}$$

$$(\text{正答率}) = \frac{TP + TN}{TP + FP + TN + FN} \qquad \text{(accuracy)}$$

- 真陽性率: 感度 (sensitivity) あるいは 再現率 (recall)
- 真陰性率: 特異度 (specificity)

4.5

F-値 (F-measure, F-score)

$$F_1 = \frac{2}{1/(\text{再現率}) + 1/(\text{適合率})} \quad (\text{調和平均})$$

$$F_\beta = \frac{\beta^2 + 1}{\beta^2/(\text{真陽性率}) + 1/(\text{適合率})} (\text{重み付き調和平均})$$

再現率(真陽性率)と適合率の調和平均

演習: さまざまな評価値

- 前回用いたデータについて，さまざまな評価値を計算してみよう

4.7

予測誤差

5.1

訓練誤差と予測誤差

- **訓練誤差** (training error): 既知データに対する誤り
- **予測誤差** (predictive error): 未知データに対する誤り
- 訓練誤差は予測誤差より良くなることが多い
  既知データの判別に特化している可能性があるため
  - 過適応 (over-fitting)
  - 過学習 (over-training)

交叉検証

- 収集したデータを訓練データと試験データに分割して用いる:
  - **訓練データ** (training data): 判別関数を構成する
  - **試験データ** (test data): 予測精度を評価する
- データの分割に依存して予測誤差の評価が偏る
- 偏りを避けるために複数回分割を行ない評価する

交叉検証法 (cross-validation; CV)

- k -重交叉検証法 (k -fold cross-validation; k -fold CV)
  - n 個のデータを k ブロックにランダムに分割
  - 第 i ブロックを除いた $k-1$ ブロックで判別関数を推定
  - 除いておいた第 i ブロックで予測誤差を評価
  - $i=1,\dots,k$ で繰り返し k 個の予測誤差で評価 (平均や分散)
- leave-one-out法 (leave-one-out CV; LOO-CV)
  - $k=n$ として上記を実行

演習: 予測誤差の評価

- 10-valid.r を確認してみよう

5.5

演習: 交叉検証による評価

- 10-cv.r を確認してみよう

5.6

演習

- 前回用いたデータについて線形・2次どちらの判別方法が望ましいか検証してみよう

5.7