判別分析

考え方

村田 昇

2020.11.13

講義の予定

- **第 1 日: 判別分析の考え方**
- 第 2 日: 分析の評価

判別分析の考え方

判別分析

- **discriminant analysis**
- 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
- 関係式: **判別関数** (discriminant function)
  - 説明変数: $X = (X_1, \ldots, X_q)$
  - 目的変数: Y ($K(\geq 2)$ 個のクラスラベル)
- 判別関数による分類:
  - 1 次式の場合: **線形判別分析** (linear discriminant analysis)
  - 2 次式の場合: **2 次判別分析** (quadratic discriminant analysis)

判別分析の例

- 検査結果から患者が病気を罹患しているか判定する
  - $X =$ 検査結果
  - $Y =$ 病気・健康
- 今日の経済指標から明日株価が上昇するか予測する
  - $X =$ 今日の経済指標
  - $Y =$ 明日株価の上昇・下降
- 今日の大気の状態から, 明日の天気を予測する
  - $X =$ 今日の大気の状態
  - $Y =$ 晴・くもり・雨・雪

判別分析の考え方

- 確率による定式化
  1. $X = \boldsymbol{x}$ の下で $Y = k$ となる **条件付確率** を計算

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

  2. 所属する確率が最も高いクラスに個体を分類
- 観測データ: n 個の $(Y, X_1, \ldots, X_q)$ の組

$$\{(y_i, x_{i1}, \ldots, x_{iq})\}_{i=1}^n$$

- 観測データから条件付確率 $p_k(\boldsymbol{x})$ を構成

条件付確率

- 以下では X は離散型の q 次元確率変数で説明
- 事象 $X = \boldsymbol{x}$ が起きたという条件の下で事象 $Y = k$ が起きる条件付確率

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

- 以下の議論は連続な確率変数でも成立

条件付確率の表現

- 条件付確率 $p_k(\boldsymbol{x})$ のモデル化の方針:
  - $p_k(\boldsymbol{x})$ を直接モデル化する (例:ロジスティック回帰)
  - $Y = k$ の下での X の条件付き確率質量関数

$$f_k(\boldsymbol{x}) = P(X = \boldsymbol{x}|Y = k) = \frac{P(X = \boldsymbol{x}, Y = k)}{P(Y = k)}$$

    のモデル化を通じて $p_k(\boldsymbol{x})$ をモデル化する
- 本講義では後者について説明

事後確率による判別

ベイズの公式

- $f_k(\boldsymbol{x})$ から $p_k(\boldsymbol{x})$ を得る数学的原理

  原因 $X = \boldsymbol{x}$ から結果 $Y = k$ が生じる確率を**結果 $Y = k$ が生じたときの原因が $X = \boldsymbol{x}$ である確率**から計算する方法
- ベイズの公式 (Bayes' formula)

$$P(Y = k|X = \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})P(Y = k)}{\sum_{l=1}^K f_l(\boldsymbol{x})P(Y = l)}$$

ベイズの公式の略証

- 定義より

$$f_k(\boldsymbol{x}) = P(X = \boldsymbol{x}|Y = k) = \frac{P(X = \boldsymbol{x}, Y = k)}{P(Y = k)}$$

- 求める条件付確率:

$$P(Y = k|X = \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})P(Y = k)}{P(X = \boldsymbol{x})}$$

- 分母の展開:

$$\begin{aligned} P(X = \boldsymbol{x}) &= \sum_{l=1}^{K} P(X = \boldsymbol{x}, Y = l) \\ &= \sum_{l=1}^{K} f_l(\boldsymbol{x})P(Y = l) \end{aligned}$$

事前確率と事後確率

- **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
- ベイズの公式による書き換え:

$$p_k(\boldsymbol{x}) = \frac{f_k(\boldsymbol{x})\pi_k}{\sum_{l=1}^{K} f_l(\boldsymbol{x})\pi_l}$$

  事前確率が特徴量の条件付確率の重みで変更される

事前確率の決め方

- 事前に特別な情報がない場合:

  データから自然に決まる確率

$$\pi_k = \frac{Y = k \text{ のサンプル数}}{\text{全サンプル数}}$$

- 事前に情報がある場合:

  例: 身長や体重などの特徴量から喫煙者か否かを判別
  - 喫煙者の身長や体重などの特徴量を収集
  - 非喫煙者の身長や体重などの特徴量を収集
  - 事前確率は (別の調査の) 日本人の喫煙者の割合を利用

線形判別分析

判別関数

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

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

  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

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

- 線形判別関数: $\boldsymbol{x}$ の 1 次式

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

同値性の確認

- 事後確率と判別関数の関係

$$\begin{aligned}
&p_k(\boldsymbol{x}) < p_k(\boldsymbol{x}) \\
&\Leftrightarrow f_k(\boldsymbol{x})\pi_k < f_l(\boldsymbol{x})\pi_l \\
&\Leftrightarrow \log f_k(\boldsymbol{x}) + \log \pi_k < \log f_l(\boldsymbol{x}) + \log \pi_l \\
&\Leftrightarrow -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^\mathsf{T}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k) + \log \pi_k \\
&\qquad < -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_l)^\mathsf{T}\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_l) + \log \pi_l \\
&\Leftrightarrow \delta_k(\boldsymbol{x}) < \delta_l(\boldsymbol{x})
\end{aligned}$$

平均・分散の推定

- 平均の推定 (クラスごとに行う)

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i:y_i=k} \boldsymbol{x}_i$$

ただし n_k は $y_i = k$ であるようなデータの総数

- 分散の推定 (まとめて行う)

$$\hat{\Sigma} = \frac{1}{n-k} \sum_{k=1}^{K} \sum_{i:y_i=k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathsf{T}}$$

R: 線形判別関数 `MASS::lda()`

- データフレームに対する分析:

```
library(MASS) # または require(MASS)
lda(formula = y の変数名 ~ x1 の変数名 + ... + xp の変数名,
    data = データフレーム)
## formula: 目的変数名 ~ 説明変数名
## data: 目的変数，説明変数を含むデータフレーム
## 書式は lm() とほぼ同じ
```

- 判別関数値の図示:

```
est <- lda(formula = y の変数名 ~ x1 の変数名 + ... + xp の変数名,
           data = データフレーム)
plot(est)
```

練習問題

- 東京の気候データを用いて以下の分析を行いなさい
  - 10 月と 11 月の気温と湿度のデータを抽出する

```
TW.data <- transform(read.csv("data/tokyo_weather_reg.csv"),
                     month=as.numeric(months(as.Date(date),
                                             abbreviate=TRUE)))
TW.subset  <- subset(TW.data,
                     subset= month %in% c(10,11),
                     select=c(temp,humid,month))
```

  - 半分のデータを用いて線形判別関数を構成し，残りのデータを用いて判別を行う

```
library(MASS)
idx <- seq(2,60,by = 2)
TW.train <- TW.subset[ idx,] # 訓練データ
TW.test  <- TW.subset[-idx,] # 試験データ
TW.lda <- lda(month ~ temp + humid, data=TW.train) # 線形判別関数の構成
TW.pred <- predict(TW.lda, newdata=TW.test) # 新しいデータの予測
```

2 次判別分析

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**

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

- 2 次判別関数: $\boldsymbol{x}$ の 2 次式

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

同値性の確認

- 事後確率と判別関数の関係

$$\begin{aligned}
& p_k(\boldsymbol{x}) < p_l(\boldsymbol{x}) \\
\Leftrightarrow\ & f_k(\boldsymbol{x})\pi_k < f_l(\boldsymbol{x})\pi_l \\
\Leftrightarrow\ & \log f_k(\boldsymbol{x}) + \log \pi_k < \log f_l(\boldsymbol{x}) + \log \pi_l \\
\Leftrightarrow\ & -\frac{1}{2}\det \Sigma_k - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\mathsf{T}\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) + \log \pi_k \\
& \quad < -\frac{1}{2}\det \Sigma_l - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_l)^\mathsf{T}\Sigma_l^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_l) + \log \pi_l \\
\Leftrightarrow\ & \delta_k(\boldsymbol{x}) < \delta_l(\boldsymbol{x})
\end{aligned}$$

平均・分散の推定

- 平均の推定 (クラスごとに行う)

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n_k}\sum_{i:y_i=k} \boldsymbol{x}_i$$

  だたし n_k は $y_i = k$ であるようなデータの総数

- 分散の推定 (クラスごとに行う)

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}\sum_{i:y_i=k} (\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^\mathsf{T}$$

R: 2 次判別関数 `MASS::qda()`

- データフレームに対する分析:

```
library(MASS) # または require(MASS)
qda(formula = y の変数名 ~ x1 の変数名 + ... + xp の変数名,
    data = データフレーム)
## formula: 目的変数名 ~ 説明変数名
## data: 目的変数，説明変数を含むデータフレーム
```

練習問題

- 東京の気候データを用いて以下の分析を行いなさい
  - 前問と同様な設定で 2 次判別を行いなさい

```
TW.qda <- qda(month ~ temp + humid, data=TW.train) # 2次判別関数の構成
TW.pred <- predict(TW.qda, newdata=TW.test) # 新しいデータの予測
```

- 別の月や変数を用いて判別分析を行いなさい

多値判別

多値判別の考え方

- 判別関数の比較 (判別関数 δ_k を比較)
- 2 値判別の統合 (組み合わせ数 ${}_nC_2$)
- $k-1$ 個の特徴量への変換 (R の線形判別)

変動の分解

- 3 種類の変動
  - $A = \sum_{i=1}^{n} (\boldsymbol{x}_i - \mu)(\boldsymbol{x}_i - \mu)^{\mathsf{T}}$: 全変動
  - $W = \sum_{i=1}^{n} (\boldsymbol{x}_i - \mu_{y_i})(\boldsymbol{x}_i - \mu_{y_i})^{\mathsf{T}}$: 群内変動
  - $B = \sum_{k=1}^{K} n_k (\mu_k - \mu)(\mu_k - \mu)^{\mathsf{T}}$: 群間変動
    (n_k はクラス k のデータ数)
- 変動の関係

$$A = W + B$$

Fisher の線形判別

- 判別のための特徴量 $Z = \alpha^{\mathsf{T}} X$
- 良い Z の基準:
  - クラス内では集まっているほど良い ($\alpha^{\mathsf{T}} W \alpha$ は小)
  - クラス間では離れているほど良い ($\alpha^{\mathsf{T}} B \alpha$ は大)
- Fisher の基準:

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

  クラス内変動を一定にしてクラス間変動を最大化する

Fisher の線形判別の解

- α は $W^{-1}B$ の固有値 (主成分分析の導出と同様)
  - $K = 2$ の場合: 最大固有値を用いる

$$\alpha \propto W^{-1}(\mu_1 - \mu_2)$$

    線形判別と一致する
  - 一般の K の場合: 第 1 から第 $K-1$ 固有値を用いる
- 判別の手続き:
  特徴量とクラスの中心までの距離を用いる
  1. $d_k = \sum_{l=1}^{K-1} (\alpha_l^{\mathsf{T}} \boldsymbol{x} - \alpha_l^{\mathsf{T}} \mu_k)^2$ を計算
  2. 最小の d_k となるクラス k に判別

練習問題

- 東京の気候データを用いて以下の分析を行いなさい
  - 9 月，10 月，11 月の気温と湿度のデータを用いて判別関数を作成しなさい.

```
TW.subset  <- subset(TW.data,
                     subset= month %in% c(9,10,11),
                     select=c(temp,humid,month))
TW.lda <- lda(month ~ temp + humid, data=TW.subset)
```

  - 別の月や変数を用いて判別分析を行いなさい