

# 判別分析

## 基本的な考え方

村田 昇

## 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する方法
- 関係式：**判別関数** (discriminant function)
  - 説明変数： $X = (X_1, \dots, X_q)$
  - 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を予測する**
  - $X$  = 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.  $X = \mathbf{x}$  の下で  $Y = k$  となる **条件付確率** を計算

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

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  $(Y, X_1, \dots, X_q)$  の組

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

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

## 条件付確率

- 以下では  $X$  は離散型の  $q$  次元確率変数として説明
- 事象  $X = \mathbf{x}$  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

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

- 連続な確率変数の場合は確率密度関数を用いる

## 条件付確率の表現

- $Y$  の条件付確率  $p_k(\mathbf{x})$  のモデル化の方針
  - $p_k(\mathbf{x})$  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

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

のモデル化を通じて  $p_k(\mathbf{x})$  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

### Bayes の公式

- $f_k(\mathbf{x})$  から  $p_k(\mathbf{x})$  を得る数学的原理  
原因  $X = \mathbf{x}$  から結果  $Y = k$  が生じる確率を結果  $Y = k$  が生じた原因が  $X = \mathbf{x}$  である確率から計算する方法
- Bayes の公式 (Bayes' formula)

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

### Bayes の公式の略証

- 定義より

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

- 求める条件付確率

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

- 分母の展開

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

## 事前確率と事後確率

- 事前確率:  $\pi_k = P(Y = k)$  (prior probability)
  - $X = \mathbf{x}$  が与えられる前に予測されるクラス確率
- 事後確率:  $p_k(\mathbf{x})$  (posterior probability)
  - $X = \mathbf{x}$  が与えられた後に予測されるクラス確率
- Bayes の公式による書き換え:

$$p_k(\mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{\sum_{l=1}^k f_l(\mathbf{x})\pi_l} = \frac{f_k(\mathbf{x})}{\sum_{l=1}^k f_l(\mathbf{x})\pi_l} \cdot \pi_k$$

- 事前確率が説明変数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  
食事・運動・飲酒・ストレスなどの生活の特徴から生活習慣病か否かを判別
  - 健常者の食事・運動・飲酒・ストレスなどの特徴量を収集
  - 罹患者の食事・運動・飲酒・ストレスなどの特徴量を収集
  - 事前確率は 別の調査の日本人の罹患率 を利用

## 線形判別分析

### 判別関数

- 判別の手続き
  1. 説明変数  $X = \mathbf{x}$  の取得
  2. 事後確率  $p_k(\mathbf{x})$  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:  $\delta_k(\mathbf{x})$  ( $k = 1, \dots, K$ )

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

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

### 線形判別

- $f_k(\mathbf{x})$  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  $\Sigma$ : すべてのクラスで共通

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

- 線形判別関数:  $\mathbf{x}$  の 1 次式

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

## 同値性の確認

- 事後確率と判別関数の関係

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

## 平均・分散の推定

- 平均の推定 (クラスごとに行う)

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

– 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)

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

## 実習

### 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 <- read.csv("data/tokyo_weather.csv")
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_est <- predict(tw_lda) # 判別関数によるクラス分類結果の取得
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

- $f_k(\mathbf{x})$  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  $\Sigma_k$ : クラスごとに異なる

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

- 2次判別関数:  $\mathbf{x}$  の2次式

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

### 同値性の確認

- 事後確率と判別関数の関係

$$\begin{aligned} p_k(\mathbf{x}) &< p_l(\mathbf{x}) \\ \Leftrightarrow f_k(\mathbf{x})\pi_k &< f_l(\mathbf{x})\pi_l \\ &\quad (\text{分母は共通}) \\ \Leftrightarrow \log f_k(\mathbf{x}) + \log \pi_k &< \log f_l(\mathbf{x}) + \log \pi_l \\ \Leftrightarrow -\frac{1}{2} \log \det \Sigma_k - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) + \log \pi_k \\ &\quad < -\frac{1}{2} \log \det \Sigma_l - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_l)^\top \Sigma_l^{-1} (\mathbf{x} - \boldsymbol{\mu}_l) + \log \pi_l \\ \Leftrightarrow \delta_k(\mathbf{x}) &< \delta_l(\mathbf{x}) \\ &\quad (\text{2次の項は右辺と左辺で共通}) \end{aligned}$$

## 平均・分散の推定

- 平均の推定 (クラスごとに行う)

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

- 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 (クラスごとに行う)

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

## 実習

### 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_est <- predict(tw_qda) # 判別関数によるクラス分類結果の取得
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

### 多値判別の構成方法

- 判別関数の比較
  - 判別関数  $\delta_k$  を比較
  - 正規分布を仮定する場合は一般には 2 次判別
- 2 値判別の統合
  - 2 クラスでの比較: 最大の組合せ数  $K C_2$
  - グループでの比較: 最大の組合せ数  $2^K - 2$
- $K-1$  個の特徴量への変換
  - 説明変数の線形結合による特徴量の構成
  - **Fisher の線形判別**

## 変動の分解

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

$$(\text{全変動}) = (\text{群内変動}) + (\text{群間変動})$$

$$A = W + B$$

## Fisher の判別分析

### Fisher の線形判別

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

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

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

- $\boldsymbol{\alpha}$  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 (線形判別と一致)

$$\boldsymbol{\alpha} \propto W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

- 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
  - 特徴量とクラスを中心までの距離を用いる
    1.  $d_k = \sum_{l=1}^{K-1} (\boldsymbol{\alpha}_l^\top \mathbf{x} - \boldsymbol{\alpha}_l^\top \boldsymbol{\mu}_k)^2$  を計算
    2. 最小の  $d_k$  となるクラス  $k$  に判別

## 演習

### 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

```
## 雨の有無を識別する例
tw_mydata <- transform(tw_data,
                       rain=factor(rain>0), # 雨の有無でラベル化する
                       month=factor(month)) # 月ごとの気候の違いの補正のため
tw_mylda <- lda(rain ~ temp + solar + wind + month,
               data=tw_mydata)
```

## 次週の予定

- 第1日：判別分析の考え方
- 第2日：分析の評価