クレジット:
UTokyo Online Education 統計データ解析 Ⅱ 2018 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 (II) 第 11 回

小池祐太

2018 年 6 月 21 日

判別分析

- **判別分析**
  - ある個体が $K(\geq 2)$ 個のクラスのいずれかに属するとき, その個体の属性 (特徴量) $X = (X_1, \ldots, X_q)$ からどのクラスに属するか予測するモデルを構築するための分析法
- 具体例
  - ある病気について, いくつかの検査結果から患者がその病気を罹患しているか否か判定する (Y: 罹患しているか否か, X: 検査結果たち)
  - 今日の経済指標から明日株価が上昇するか否か予測する (Y: 明日の株価は上昇するか下降するか, X: 今日の経済指標たち)
  - 今日の大気の状態から, 明日の天気が晴・くもり・雨・雪のいずれであるかを予測する (Y: 明日の天気, X: 今日の大気の状態を表す指標)

判別分析

- 数学的には, クラスラベルを表す質的変数を $Y \in \{1, \ldots, K\}$ としたとき, $X = \boldsymbol{x}$ の下で $Y = k$ となる条件付き確率

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

  に対するモデルを構築することが目的となる
  - 所属する確率が最も高いクラスラベルに個体を分類すればよい (状況によっては他の分類基準も適用可能)
  - ただし, 直接判別基準を構築するアプローチもある (例: サポートベクターマシン)
- 観測データとしては, 組 $(Y, X_1, \ldots, X_q)$ に対する n 個の観測データ

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

  が与えられている状況を考える

判別分析

- 以下しばらくの間 X が離散型の q 次元確率変数である場合を考える
- この場合,

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

である (“事象 $X = \boldsymbol{x}$ が起きた下で事象 $Y = k$ が起きる条件付き確率”)

判別分析

- $p_k(\boldsymbol{x})$ をモデル化するアプローチとしては以下の 2 通りの方法がある:
  1. $p_k(\boldsymbol{x})$ を直接モデル化する (例: ロジスティック回帰).
  2. $Y = k$ 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.
- 本講義では後者のアプローチについて説明する

ベイズの公式

- $f_k(\boldsymbol{x})$ のモデル化を通じて $p_k(\boldsymbol{x})$ のモデルが得られることの数学的原理は, 次の**ベイズの公式 (Bayes' formula)** によって与えられる:

定理 1 (ベイズの公式)

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

- 「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, 原因が $X = \boldsymbol{x}$ だった確率」から計算する方法を与える
  - 「今日の天気が晴・くもり・雨・雪のそれぞれだったときの昨日の大気の状態」を調べることで, ある日の大気の状態からその次の日の天気が晴・くもり・雨・雪のそれぞれになる確率を計算する

- ベイズの公式は以下のようにして証明できる
- まず, 定義より

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

であるから,

$$P(X = \boldsymbol{x}, Y = k) = f_k(\boldsymbol{x})P(Y = k) \tag{1}$$

が成り立つ.

- これを $P(Y = k|X = \boldsymbol{x})$ の定義式に代入して

$$P(Y = k|X = \boldsymbol{x}) = \frac{f_k(\boldsymbol{x})P(Y = k)}{P(X = \boldsymbol{x})} \tag{2}$$

を得る.

- 一方で, Y は $1, \ldots, K$ のうちいずれか一つの値のみ取ることに注意すると,

$$P(X = \boldsymbol{x}) = \sum_{l=1}^{K} P(X = \boldsymbol{x}, Y = l)$$

が成り立つ

- 上式右辺の総和の各項に (1) 式を適用して

$$P(X = \boldsymbol{x}) = \sum_{l=1}^{K} f_l(\boldsymbol{x}) P(Y = l) \tag{3}$$

を得る

- この式を (2) 式に代入することで, 証明すべき等式が得られる

ベイズの公式

- $Y = k$ となる確率を $\pi_k = P(Y = k)$ と書くことにすると, ベイズの公式より,

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

  が成り立つ
- 従って, $\pi_1, \ldots, \pi_K$ がわかっている, もしくはデータから推定できるのであれば, $f_k(\boldsymbol{x})$ をモデル化することで $p_k(\boldsymbol{x})$ のモデルが得られる
- $\pi_1, \ldots, \pi_K$ は**事前確率 (prior probability)** と呼ばれ, 特徴量が与えられる前に予測できるそれぞれのクラスに属する確率である
  - 対応して, $p_k(\boldsymbol{x})$ は**事後確率 (posterior probability)** と呼ぶことがある ($X = \boldsymbol{x}$ であることがわかる前と後という意味で)

ベイズの公式

- 事前確率に関する特別な情報がない場合は, π_k はデータから自然に決まる確率

$$\frac{Y = k \text{ であるサンプル数}}{\text{全サンプル数}}$$

  で推定される
- 一方で, 例えば日本人のサンプルから身長や体重などの特徴量を観測したデータから, その人が喫煙者か否かを判別するためのモデルを構築するといった状況の場合, 事前確率として日本人の喫煙者の割合といったデータを使うことも考えられる
  - $Y = 1$ が喫煙者を表し, $Y = 2$ が非喫煙者を表す場合, π_1 として日本人の喫煙者の割合を使い, π_2 として日本人の非喫煙者の割合を使うということが考えられる

ベイズの公式

- 以上の議論は X が連続型の場合でも, $f_k(\boldsymbol{x})$ を $Y = k$ の下での X の条件付き確率密度関数とすればそのまま成立する
- すなわち, この場合 $f_k(\boldsymbol{x})$ はクラス k に属するようなサンプルの場合に X が従う確率分布の確率密度関数を表す
- なお, この場合, 条件付き確率 $p_k(\boldsymbol{x}) = P(Y = k|X = \boldsymbol{x})$ は「X が $\boldsymbol{x}$ に非常に近いという条件の下で $Y = k$ となる確率」を意味する
  - この場合常に $P(X = \boldsymbol{x}) = 0$ なので, 先ほどの定義は意味をなさないため
  - 「非常に近い」という部分は, 数学的には極限操作によって定義する. 配布資料参照のこと

判別関数

- 実際に観測データに基づいて, 特徴量が $X = \boldsymbol{x}$ であるようなデータの属するクラスを判別する際には, $p_k(\boldsymbol{x})$ を最大にするようなクラス k にデータを分類する
- 従って, 関数 $\delta_k(\boldsymbol{x})$ $(k = 1, \ldots, K)$ で,

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

を満たすようなものが存在すれば, $\delta_k(\boldsymbol{x})$ を最大化するようなクラス k にそのデータを分類すればよいことになる

  - このような関数 $\delta_k(\boldsymbol{x})$ を**判別関数**と呼ぶ

線形判別分析

- **線形判別分析 (linear discriminant analysis)** では, $f_k(\boldsymbol{x})$ をクラスごとに異なる平均ベクトル $\boldsymbol{\mu}_k$ をもつが, すべてのクラスで共通の共分散行列 Σ をもつような q 変量正規分布の確率密度関数としてモデル化する:

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

- このモデル化の下, **線形判別関数 (linear discriminant function)**

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

を最大化するようなクラス k にデータを分類すればよいという結論が導かれる

実際,

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

が成り立つ

線形判別分析

- $K = 2$ の場合, 方程式

$$\delta_2(\boldsymbol{x}) - \delta_1(\boldsymbol{x}) = 0$$

で定まる q 次元空間内の (超) 平面を考え, そのデータの特徴量がその平面の下側であれば $Y = 1$ と判別し, 上側であれば $Y = 2$ と判別すればよい ($q = 2$ の場合は直線)

- 簡単な計算によって

$$\delta_2(\boldsymbol{x}) - \delta_1(\boldsymbol{x}) = \left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) \cdot \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

と書き直せることがわかるので, 上述の平面は点 $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ を通り法線ベクトルが $\Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ の平面である

線形判別分析

- 線形判別関数の計算のためには各クラスごとの特徴量の平均ベクトル $\boldsymbol{\mu}_k$ およびすべてのクラスで共通の特徴量の共分散行列 Σ を計算する必要があるが, これらはそれぞれ

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

で推定すればよい. ここに, n_k は $y_i = k$ であるようなデータの総数を表す

線形判別分析: R での実行

- パッケージ MASS には線形判別分析を実行するための関数 lda( ) が用意されている. 書式は関数 lm( ) とほとんど同じである (クラスラベルを目的変数, 特徴量を説明変数とする)
- 実行例 lda2g.r, tenki.r, lda.r

2 次判別分析

- **2 次判別分析 (quadratic discriminant analysis)** では, $f_k(\boldsymbol{x})$ をクラスごとに異なる平均ベクトル $\boldsymbol{\mu}_k$ と共分散行列 Σ_k をもつような q 変量正規分布の確率密度関数としてモデル化する:

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

- いま,

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

が成り立つから, **2 次判別関数**

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

を最大化するようなクラス k にデータを分類すればよい

2 次判別分析

- 2 次判別関数の計算のためには各クラスごとの特徴量の平均ベクトル $\boldsymbol{\mu}_k$ および共分散行列 Σ_k を計算する必要があるが, これらはそれぞれ

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

で推定すればよい. ここに, n_k は $y_i = k$ であるようなデータの総数を表す

2 次判別分析: R での実行

- パッケージ MASS には 2 次判別分析を実行するための関数 qda( ) が用意されている. 書式は関数 lda( )(従って関数 lm( )) とほとんど同じである (クラスラベルを目的変数, 特徴量を説明変数とする)
- 実行例 qda.r