Diskriminační analýza

Diskriminační analýza

Data:

- $(x_{i,1}, \ldots, x_{i,p}, Y_i)^T$ pro $i = 1, \ldots, n$.
- $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^T$ je vektor (spojitých) regresorů .
- Y_i udává příslušnost i-tého pozorování k dané skupině.
- Y_i je kategoriální proměnná nabývající hodnot $1, 2, \ldots, J$.

Cíl: Na základě dat zkonstruovat rozhodovací pravidlo, které bude co nejlépe klasifikovat nová pozorování $(x^*_{i,1}, \ldots, x^*_{i,p})^T$ do příslušné skupiny.

Metody diskriminační analýzy

- Úloha supervised learning (učení s učitelem).
- Zakladatel: R. A. Fisher (1936) – klasifikace kosatců (iris).
- Způsoby odvození klasifikačního pravidla:
  - Kanonická diskriminační analýza (kombinace PCA a MANOVA - hledání nového souřadnicového systému, který maximalizuje podíl vnitro a meziskupinové variability).
  - Parametrické metody (lineární a kvadratická diskriminační analýza).
  - Neparametrické metody (k-nearest neighbors, metody založené na jádrových odhadech hustoty, na hloubce dat, apod.)

Parametrické metody - rozhodovací pravidla

- Nechť jsou naše data generována náhodným vektorem $\mathbf{X} = (X_1, \ldots, X_p)'$ a nechť Y značí náhodnou veličinu udávající příslušnost daného pozorování k dané skupině; Y nabývá hodnot $1, 2, \ldots, J$.
- Předpokládejme, že známe hustotu rozdělení náhodného vektoru $\mathbf{X}$ pro j-tou skupinu, tj. nechť rozdělení $\mathbf{X}|Y = j$ má hustotu $p_j(\mathbf{x})$ (známá p-rozměrná hustota).
- V praxi se pro určení pravidel nejčastěji používá princip maximální věrohodnosti a Bayesovský přístup.
- Princip maximální věrohodnosti: Pozorování $\mathbf{x}^*$ zařaď do skupiny $\arg\max\{p_j(\mathbf{x}^*); j = 1, \ldots, J\}$.
- Bayesovský přístup: Pozorování $\mathbf{x}^*$ zařaď do skupiny $\arg\max\{p_j(\mathbf{x}^*)\pi_j; j = 1, \ldots, J\}$, kde π_j je apriorní pravděpodobnost, že pozorování patří do skupiny j, tj. $\pi_j = P(Y = j)$.
- Bayesovské pravidlo minimalizuje pravděpodobnost špatné klasifikace.

Lineární diskriminační analýza

- Předpokládejme, že $p_j(\mathbf{x})$ jsou hustoty p-rozměrného normálního rozdělení $\mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ se stejnými variančními maticemi $\boldsymbol{\Sigma}$.
-

$$p_j(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}, \quad \mathbf{x} \in \mathbb{R}^p.$$

- Bayesovské pravidlo - můžeme hledat maximum $\log p_j(\mathbf{x})\pi_j = \log p_j(\mathbf{x}) + \log \pi_j$.
- $\log p_j(\mathbf{x}) + \log \pi_j = \log \pi_j + \boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \mathbf{x} - 1/2 \boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + K$.
- $L_j(\mathbf{x}) = \log p_j(\mathbf{x}) + \log \pi_j = \log \pi_j + \boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \mathbf{x} - 1/2 \boldsymbol{\mu}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j$ se nazývá lineární diskriminační funkce - tu maximalizujeme přes $j = 1, \ldots, J$.
- $K = -p/2 \log 2\pi - 1/2 \log |\boldsymbol{\Sigma}| - 1/2 \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x}$ nezávisí na j.

Lineární diskriminační analýza - praktické použití

- Na trénovacích datech nejprve odhadneme neznámé parametry modelu a poté model vyzkoušíme na testovacích datech.
- π_j neznáme, musíme je odhadnout. Pokud věříme proporcionálnímu zastoupení skupin v našich datech, pak $\widehat{\pi}_j = \frac{n_j}{n}$ je relativní četnost j-té skupiny v datech (n_j je počet pozorování z j-té skupiny). Jinak využijeme princip neurčitosti $\widehat{\pi}_j = \frac{1}{J}$ - vede na princip maximální věrohodnosti.
- $\boldsymbol{\mu}_j$ odhadneme pomocí výběrového průměru v j-té skupině.
- Označme $\mathbf{S}_j$ odhad varianční matice v j-té skupině, pak společný odhad společné varianční matice je $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-J}\left((n_1 - 1)\mathbf{S}_1 + \ldots + (n_J - 1)\mathbf{S}_J\right)$
- Pro klasifikaci použijeme výběrovou verzi $L_j(\mathbf{x})$, kde $\pi_j, \boldsymbol{\mu}_j$ a $\boldsymbol{\Sigma}$ nahradíme jejich odhady.

Kvadratická diskriminační analýza

- Předpokládejme, že $p_j(\mathbf{x})$ jsou hustoty p-rozměrného normálního rozdělení $\mathcal{N}_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$.
-

$$p_j(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_j|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) \right\}, \quad \mathbf{x} \in \mathbb{R}^p.$$

- $\log p_j(\mathbf{x}) + \log \pi_j =$
  $\log \pi_j - 1/2 \log |\boldsymbol{\Sigma}_j| + \boldsymbol{\mu}_j' \boldsymbol{\Sigma}_j^{-1} \mathbf{x} - 1/2 \boldsymbol{\mu}_j' \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - 1/2 \mathbf{x}' \boldsymbol{\Sigma}_j^{-1} \mathbf{x} + K.$
- $Q_j(\mathbf{x}) = \log \pi_j - 1/2 \log |\boldsymbol{\Sigma}_j| + \boldsymbol{\mu}_j' \boldsymbol{\Sigma}_j^{-1} \mathbf{x} - 1/2 \boldsymbol{\mu}_j' \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j - 1/2 \mathbf{x}' \boldsymbol{\Sigma}_j^{-1} \mathbf{x}$ se nazývá kvadratická diskriminační funkce - tu maximalizujeme přes $j = 1, \ldots, J$.
- $K = -p/2 \log 2\pi$ nezávisí na j.

Kvadratická diskriminační analýza - praktické použití

- Na trénovacích datech nejprve odhadneme neznámé parametry modelu a poté model vyzkoušíme na testovacích datech.
- Pro klasifikaci použijeme výběrovou verzi $Q_j(\mathbf{x})$, kde $\pi_j, \boldsymbol{\mu}_j$ a $\boldsymbol{\Sigma}_j$ nahradíme jejich odhady.

Neparametrická diskriminační analýza

- Idea: hustotu $p_j(\mathbf{x})$ pro j-tou skupinu neznáme - odhadneme ji pomocí trénovacích dat.
- Typicky se uvažují jádrové odhady hustot $p_j(\mathbf{x})$, označme je $\widehat{p_j}(\mathbf{x})$.
- Upravíme Bayesovské rozhodovací pravidlo: Pozorování $\mathbf{x}^*$ zařaď do skupiny $\arg\max\{\widehat{p_j}(\mathbf{x}^*)\pi_j; j = 1, \ldots, J\}$.
- Výpočetně náročné (obzvláště ve vyšších dimenzích).
- Ve vyšších dimenzích trpí prokletím dimenzionality (odhady hustot mají obrovskou variabilitu - jsou nepřesné).

Metoda k nejbližších sousedů

- k nearest neighbors (kNN).
- Zvolíme nějakou vzdálenost (např. Mahalanobisovu) v $\mathbb{R}^p$.
- Pro dané pozorování $\mathbf{x}^*$ najdeme jeho k nejbližších sousedů (v této vzdálenosti).
- Upravíme Bayesovské rozhodovací pravidlo, $p_j(\mathbf{x}^*)$ odhadneme pomocí "histogramu" na základě k nejbližších sousedů.
- $\widehat{p_j}(\mathbf{x}^*) = \frac{k_j}{n_j}$, kde k_j je počet pozorování ze skupiny j mezi k nejbližšími sousedy $\mathbf{x}^*$.
- Pozorování $\mathbf{x}^*$ zařaď do skupiny
  $\arg\max\{\widehat{p_j}(\mathbf{x}^*)\pi_j; j = 1, \ldots, J\} = \arg\max\left\{\frac{k_j}{n_j}\pi_j; j = 1, \ldots, J\right\}$.
- Odhadneme-li π_j pomocí relativních četností, pak $\mathbf{x}^*$ zařaď do skupiny $\arg\max\{k_j; j = 1, \ldots, J\}$. Tedy $\mathbf{x}^*$ zařaď do skupiny, která je nejčastěji zastoupena mezi k nejbližšími sousedy $\mathbf{x}^*$.

Metody založené na hloubce dat

- Maximal depth classifier.
- Pozorování $\mathbf{x}^*$ zařaď do skupiny $\arg\max\{D(\mathbf{x}^*, P_j); j = 1, \ldots, J\}$.
- $D(\mathbf{x}^*, P_j)$ je hloubka bodu $\mathbf{x}^*$ vzhledem k rozdělení j-té skupiny P_j.