Moderní metody diskriminační analýzy

3) Neparametrická diskriminační analýza

sdružené hustoty v j-té skupině $p_j(x)$ obecně neznáme → odhadneme ji z našich (trénovacích) dat $\rightsquigarrow$ $\widehat{p_j(x)}$ – jádrový odhad p-rozměrné hustoty

- Neznámé hustoty jsou p-rozměrné → problém ve více dimenzích (prokletí dimenzionality

Bayesovské rozhodovací pravidlo: x^* zařaď do třídy $\underset{j=1,\dots,J}{\arg\max}\ \hat{\pi}_j \cdot \widehat{p_j(x^*)}$.

4) Metoda k nejbližších sousedů (kNN)

– zvolíme vhodnou metriku a počet nejbližších sousedů $k \in \mathbb{N}$

Bayesovské pravidlo: $\underset{j=1,\dots,J}{\arg\max}\ \hat{\pi}_j \cdot \widehat{p_j(x^*)} = \underset{j=1,\dots,J}{\arg\max}\ \frac{\cancel{n_j}}{\underset{\text{konst.}}{n}} \cdot \frac{k_j}{\cancel{n_j}} = \underset{j=1,\dots,J}{\arg\max}\ k_j$

$\hat{\pi}_j = \frac{n_j}{n}$, $\widehat{p_j(x^*)} = \frac{k_j}{n_j}$

k_j ← počet pozorování ze třídy j mezi k nejbližšími sousedy x^*

n_j ↑ počet pozorování ze třídy j

5.) Diskriminační analýza založená na hloubce dat

– hloubka dat = zobecnění kvantilu a pořadí ve vyšších dimenzích

- Poloprostorová (Tukeyho) hloubka bodu x vzhledem k dat. souboru X: $d(x, X) = \min_{\|u\|=1} \frac{1}{n} \left|\{ i : x_i^T u \le x^T u,\ i = 1,\dots,n \}\right|$

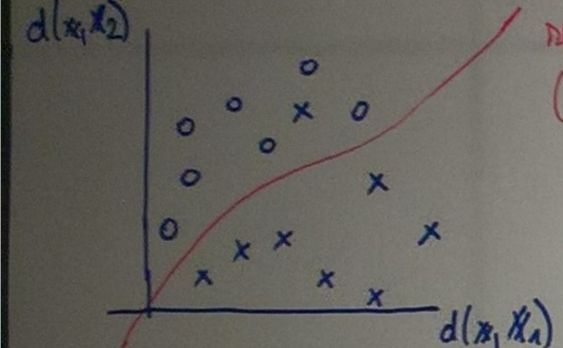
- Mahalanobisova hloubka: $d(x, X) = \dfrac{1}{1 + (x-\hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu})}$, kde $\hat{\mu}$ je výběrový průměr, $\hat{\Sigma}$ je výběrová kovarianční matice

Klasifikátor maximální hloubky:

x^* zařaď do třídy $\underset{j=1,\dots,J}{\operatorname{argmax}} \; \pi_j \cdot d(x^*, X_j)$, kde X_j je datová matice odpovídající pozorováním ze skupiny j.

mějme $J = 2$ (2 skupiny)

DD plot je 2-rozměrný graf $[d(x, X_1), d(x, X_2)]$ pro všechny body x z datového souboru



separátor, který nejlépe odděluje obě skupiny
(má minimální error rate)
- přímka
- polynom

DDα-klasifikátor: rozšíříme hloubkový prostor pro polynomy do řádu r („berou se jen ty členy, které vedou k minimalizaci pravd. chyby klasifikace)

Jako klasifikátor se vezme lineární kombinace výše uvedených prvků.

outsider = bod, který má nulovou hloubku vůči oběma skupinám (speciální zacházení)

Pro více skupin ($J>2$): uvažujeme všechny dvojice a vybereme většinový výsledek

6.) DA směsí (mixture DA)

$P_j(x)$ je směs normálních rozdělení $\quad P_j(x)$ je hustota $\displaystyle\sum_{r=1}^{K_j} \overset{\text{váhy směsí}}{\pi_{jr}} \cdot \underbrace{N_p(\mu_{jr}, \Sigma)}_{\text{prototyp}}$

$0 < \pi_{jr} < 1 \quad \pi_{j1} + \pi_{j2} + \dots + \pi_{jK_j} = 1 \quad \forall j$

K_j = počet prototypů (známé)

Bayesovské rozhodovací pravidlo – EM algoritmus

7.) Regularizovaná DA

- kompromis mezi LDA a QDA

$p_j(x)$ je hustota $N_p(\hat{\mu}_j, \hat{\Sigma}_j(\lambda))$

$$\hat{\Sigma}_j(\lambda) = (1-\lambda)\hat{\Sigma}_j + \lambda \cdot \hat{\Sigma}$$

označme $\hat{\Sigma}_1, \dots, \hat{\Sigma}_J$ výběrové kovarianční matice

$$\hat{\Sigma} = \frac{1}{n-J}\left((n_1-1)\hat{\Sigma}_1 + (n_2-1)\hat{\Sigma}_2 + \dots + (n_J-1)\hat{\Sigma}_J\right)$$

případně navíc $\hat{\Sigma}_j(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}_j(\lambda) + \frac{\gamma}{p} \cdot tr\left(\hat{\Sigma}_j(\lambda)\right) \cdot \mathbb{I}_p$.