Analýza a klasifikace dat
Bayesov klasifikátor

Institut biostatistiky a analýz
Masarykova univerzita

7. října 2012

Bayesov vzorec

- Bayesov vzorec

$$P(\omega_i|\mathbf{x}) = \frac{\mathbf{p(x|\omega_i)P(\omega_i)}}{\mathbf{p(x)}} = \frac{\mathbf{f_i(x)\pi_i}}{\sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{f_j(x)\pi_j}}$$

  - skupiny $\omega_1, \omega_2, \ldots, \omega_k$
  - π_i - apriórna pravdepodobnosť skupiny ω_i
  - f_i - hustota skupiny ω_i
- Kritéria klasifikácie
  - minimalizovať očakávanú cenu za chybnú klasifikáciu (minimalizovať strednú stratu)
  - minimalizovať celkovú pravdepodobnosť chybného zaradenia
  - maximalizovať aposteriórne pravdepodobnosti
  - maximalizovať pravdepodobnosti

Minimalizovanie strednej straty

- Označme R_i ako množinu tých výsledkov, ktoré zaradíme do skupiny ω_i
- Označme $c(\omega_i|\omega_j)$ ako cenu, ktorú zaplatíme, keď prvok zo skupiny ω_j nesprávne zaradíme do skupiny ω_i, $c(\omega_i|\omega_i) = 0$
- Podmienené očakávanie ceny za chybné zaradenie pre skupinu ω_i

$$ECM(i) = \sum_{j=1}^{k} c(\omega_j|\omega_i) P(\mathbf{X} \in R_j | \mathbf{X} \in \omega_i)$$

- Očakávaná cena chybného zaradenia je

$$ECM = \sum_{i=1}^{k} \pi_i ECM(i) = \sum_{i=1}^{k} \pi_i \left(\sum_{j=1}^{k} c(\omega_j|\omega_i) P(\mathbf{X} \in R_j | \mathbf{X} \in \omega_i) \right)$$

Minimalizovanie strednej straty

- Zaraďovacie pravidlo
  - $\mathbf{x}_0$ zaradíme do skupiny ω_i, $i = 1, 2, \ldots, k$, pre ktorú bude mať funkcia $\mathbf{g}_i(\mathbf{x}_0)$ najmenšiu hodnotu

$$\mathbf{g}_i(\mathbf{x}_0) = \sum_{j=1}^{k} \pi_j c(\omega_i|\omega_j) f_j(\mathbf{x}_0)$$

Minimalizovanie celkovej pravdepodobnosti chybného zaradenia

- V tomto prípade berieme cenu za chybné zaradenie rovnakú pre všetky skupiny, $c(\omega_2|\omega_1) = c(\omega_3|\omega_1) = \cdots = c(\omega_{k-1}|\omega_k) = 1$
- Ale $c(\omega_i|\omega_i) = 0$
- Zaraďovacie pravidlo
  - $\mathbf{x}_0$ zaradíme do skupiny ω_i, $i = 1, 2, \ldots, k$, pre ktorú bude mať funkcia $\mathbf{g}_i(\mathbf{x}_0)$ najmenšiu hodnotu

$$\mathbf{g}_i(\mathbf{x}_0) = \sum_{j=1, j\neq i}^{k} \pi_j f_j(\mathbf{x}_0)$$

Maximalizovanie aposteriórnych pravdepodobností

- Zaraďovacie pravidlo
  - $\mathbf{x}_0$ zaradíme do skupiny ω_i, $i = 1, 2, \ldots, k$, pre ktorú bude mať funkcia $\mathbf{g}_i(\mathbf{x}_0)$ najväčšiu hodnotu

$$\mathbf{g}_i(\mathbf{x}_0) = f_i(\mathbf{x}_0)$$

Maximalizovanie pravdepodobnosti

- Nepoznáme apriórne pravdepodobnosti, preto ich zvolíme rovnaké $\pi_i = \frac{1}{k}$
- Ďalší výpočet je rovnaký ako pre minimalizovanie strednej straty

Hodnotenie úspešnosti klasifikácie

- Keby sme prvok zaraďovali náhodne len na základe apriórnych pravdepodobností, celková pravdepodobnosť mylnej klasifikácie by bola

$$p = \sum_{i=1}^{k} \pi_i (1 - \pi_i)$$

- pre $k = 2$ je $p = 0,5$, pre $k = 3$ je $p = 0,67$
- Využitie informácie obsiahnutej v dátach a použitie vhodného zaraďovacieho kritéria by malo túto pravdepodobnosť chybného zaradenia podstate znížiť
- Pravdepodobnosť chybnej klasifikácie je preto užitočnou informáciou o kvalite zaraďovacieho kritéria

Resubstitúcia

- Najjednoduchší odhad chybnej klasifikácie
- Zaraďovanie kritéria použijeme na dáta, z ktorých sme ich získali
- Vedie k podhodnoteniu odhadovaných pravdepodobností
- Ak kritérium nedosahuje dobré výsledky na dátach, z ktorých bolo odvodené, môžeme očakávať, že u nových dát bude pracovať ešte horšie

Cross-validation

- Rozdelenie súboru na dve skupiny
  - použitím dát jednej skupiny určíme zaraďovacie kritéria
  - dáta druhej skupiny klasifikujeme pomocou týchto odvodených kritérií a porovnáme so skutočným zaradením do jednotlivých skupín
- Dostaneme nestranný odhad pravdepodobnosti mylnej klasifikácie
- Nevýhodou je, že množstvo dát, ktoré máme k dispozícii musí byť dostatočne veľké, lebo časť z neho nepoužijeme na určenie klasifikačného kritéria
- Takto odhadnuté kritéria budú horšie, ako keby sme na ich určenie použili celý súbor dát

Krížové overovanie: "Jackknife procedure"

- Kritérium je odhadnuté na základe údajov o všetkých prvkoch okrem i-tého, $i = 1, 2, \ldots, n$
- Následne je i-tý prvok zaradený pomocou tohto kritéria a toto zaradenie je porovnané so skutočným
- Odhad je takmer nestranný

Pravdepodobnosť chybnej klasifikácie

- Pravdepodobnosť chybnej klasifikácie odhadujeme ako pomer chybne zaradených prvkov ku celkovému počtu prvkov

$$p = \sum_{i=1}^{n} \sum_{\hat{i}=1}^{n} \frac{n_{i\hat{i}}}{n}, \qquad i \neq \hat{i}$$

- Konfusná matica
  - matica typu *kxk*
  - na diagonále má správne zaradené prvky, mimo nesprávne zaradené s ohľadom na klasifikáciu

- Upravenie súboru, odstránenie rušenia,...
- Grafické zobrazenie dát
- Určenie rozloženia pravdepodobnosti jednotlivých skupín, testy dobrej zhody pre jednotlivé rozloženia
- Určenie parametrov rozložení a mnohorozmerného rozloženia pre jednotlivé skupiny
- Výber vhodného klasifikačného kritéria
- Úspešnosť klasifikácie
- Porovnanie s iným možným použiteľným klasifikačným kritériom