

Analýza a klasifikace dat

Bayesov klasifikátor

Institut biostatistiky a analýz
Masarykova univerzita

4. listopadu 2012

Bayesov vzorec

- Bayesov vzorec

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

- skupiny $\omega_1, \omega_2, \dots, \omega_k$
 - 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 $c(\omega_j|\omega_i)$ 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

- Zaradovacie pravidlo
 - \mathbf{x}_0 zaradíme do skupiny ω_i , $i = 1, 2, \dots, k$, 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, $c(\omega_2|\omega_1) = c(\omega_3|\omega_1) = \dots = c(\omega_{k-1}|\omega_k) = 1$
- Ale $c(\omega_i|\omega_i) = 0$
- Zaraďovacie pravidlo
 - \mathbf{x}_0 zaradíme do skupiny ω_i , $i = 1, 2, \dots, k$, 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 , $i = 1, 2, \dots, k$, 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 zarad'ovali 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 zarad'ovacieho kritéria by malo túto pravdepodobnosť chybného zaradenia podstate znížiť
- Pravdepodobnosť chybnnej klasifikácie je preto užitočnou informáciou o kvalite zarad'ovacieho kritéria

Resubstitúcia

- Najjednoduchší odhad chybnnej klasifikácie
- Zařadovanie kritéria použijeme na dáta, z ktorých sme ich získali
- Vede 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 zarad'ovacie 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

Křížové overovanie: "Jackknife procedure"

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

Pravdepodobnosť chybnjej klasifikácie

- Pravdepodobnosť chybnjej klasifikácie odhadujeme ako pomer chybnne zaradených prvkov ku celkovému počtu prvkov

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

- Konfuzná matica
 - matica typu $k \times k$
 - 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

Analýza a klasifikace dat

Lineárna klasifikácia

Institut biostatistiky a analýz
Masarykova univerzita

4. listopadu 2012

Lineárna diskriminácia

- Úlohou je nájsť pre skupiny prípadov reprezentovaných ako n -rozmerný vektor lineárnu diskriminačnú funkciu v tvare

$$g(\mathbf{x}) = a_0 + a_1x_1 + \cdots + a_nx_n,$$

- a_0 je prah diskriminačnej funkcie, konštanta
- $a_i, i = 1, 2, \dots, n$, sú váhové koeficienty pre danú skupinu ω_i

Lineárna diskriminácia - dichotomická úloha

- Máme dve skupiny ω_1 a ω_2
- Diskriminačnú funkciu môžeme napísať v tvare

$$g(\mathbf{x}) = \mathbf{w}'\mathbf{x} + w_0$$

- pozorovanie \mathbf{x} zaradíme do skupiny ω_1 ak $y(\mathbf{x}) \geq 0$
- pozorovanie \mathbf{x} zaradíme do skupiny ω_2 ak $y(\mathbf{x}) < 0$
- hraničná priamka:

$$y(\mathbf{x}) = \mathbf{w}'\mathbf{x} + w_0 = 0$$

- \mathbf{w} - normálový vektor hraničnej priamky

Lineárna diskriminácia pre viac tried

- Zařadovacie pravidlo je v tvare

$$g_r(\mathbf{x}) = \mathbf{w}'_r \mathbf{x} + w_0$$

- $\mathbf{x}_0 \in \omega_i : g_i(\mathbf{x}) > g_j(\mathbf{x}), \quad i \neq j$
- určuje sa hraničná priamka pre každú dvojicu skupín

Metoda najmenších štvorcov

- Označme pre k skupín a n pozorovaní
 - $\tilde{\mathbf{x}} = (1, \mathbf{x}')'$
 - $\tilde{\mathbf{w}} = (w_0, \mathbf{w}')'$
 - $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \dots, \tilde{\mathbf{w}}_k)$
 - $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)$
 - \mathbf{T} - matica vyjadrujúca príslušnosť ku skupine
- koeficienty $\tilde{\mathbf{w}}$ určíme pomocou metódy najmenších štvorcov

$$\tilde{\mathbf{W}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{T}$$

- pozorovanie zaradíme do tej skupiny, pre ktorú má zaradovacie pravidlo najväčšiu hodnotu

$$g_i(\mathbf{x}) = \tilde{\mathbf{w}}_i'\tilde{\mathbf{x}}$$

Fisherova diskriminačná funkcia

- Úlohou je nájsť takú lineárnu kombináciu sledovaných premenných $Y = \mathbf{v}'\mathbf{x}$, aby lepšie ako ktorákoľvek iná lineárna kombinácia separovala skupiny v tom zmysle, že jej vnútroskupinová variabilita bude čo najmenšia a medziskupinová variabilita čo najväčšia.

Fisherova diskriminácia - dichotomická úloha

- Zaradovacie pravidlo pre dichotomickú úlohu

$$g(\mathbf{x}) = \frac{\mathbf{v}'\mu_1 - \mathbf{v}'\mu_2}{\mathbf{v}'\Sigma\mathbf{v}}$$

- chceme maximalizovať
- koeficienty vyrátame

$$\mathbf{v} = \Sigma^{-1}(\mu_1 - \mu_2)$$

- potom

$$Y = \mathbf{x}'\mathbf{v} = \mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_2)$$

- stred medzi skupinami sa určí pomocou vzťahu

$$c = \frac{1}{2}(\mu_1'\mathbf{v} + \mu_2'\mathbf{v}) = \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

- zaradíme do skupiny ω_1 ak $\mathbf{x}'\mathbf{v} > c$

Fisherova diskriminácia

- Z predpokladu viacrozmerného normálneho rozloženia pozorovaní v jednotlivých skupinách sa dá odvodiť zařadovacie pravidlo, ktoré zařadí pozorovanie \mathbf{x}_0 do skupiny ω_1 (v prípade dvoch skupín) ak

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} > 1$$

Fisherova diskriminácia - Bayesov prístup

- Bayesov vzorec pre dve skupiny

$$\frac{\pi_i f_i(\mathbf{x})}{\pi_1 f_1(\mathbf{x}) + \pi_2 f_2(\mathbf{x})}$$

- V prípade zařadovania podľa maximálnej aposteriórnej pravdepodobnosti zařadíme pozorovanie \mathbf{x}_0 do skupiny ω_1 ak

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} > \frac{\pi_2}{\pi_1}$$

- v prípade, že berieme do úvahy aj nejakú strátovú funkciu, môžeme tvar upraviť na

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} > \frac{c_2 \pi_2}{c_1 \pi_1}$$

- Z normálneho rozloženia môžeme odvodiť zaraďovaciu funkciu

$$g_i(\mathbf{x}) = \mu_i' \Sigma^{-1} \mathbf{x} - \frac{1}{2} \mu_i' \Sigma^{-1} \mu_i + \ln(\pi_1)$$

- odhady
 - odhadom strednej hodnoty μ_i je výberový priemer ($\bar{\mathbf{x}}$)
 - odhadom variačnej matice pre k je spoločná variančná matica $\mathbf{S} = \frac{1}{k} \sum_{i=1}^k \mathbf{S}_i$