Analýza a klasifikace dat
Bayesov klasifikátor

Institut biostatistiky a analýz
Masarykova univerzita

17. listopadu 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

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

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

Analýza a klasifikace dat
Lineárna klasifikácia

Institut biostatistiky a analýz
Masarykova univerzita

17. listopadu 2012

Lineárna diskriminácia

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

$$g(\mathbf{x}) = a_0 + a_1 x_1 + \cdots + a_n x_n,$$

  - a_0 je prah diskriminačnej funkcie, konštanta
  - $a_i, i = 1, 2, \ldots, n$, 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

- Zaraďovacie pravidlo je v tvare

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

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

Metoda najmenších štvorcov

- Označme pre k skupín a n pozorovaní
  - $\tilde{\mathbf{x}} = (1, \mathbf{x}')'$
  - $\tilde{\mathbf{w}} = (w_0, \mathbf{w}')'$
  - $\tilde{\mathbf{W}} = (\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_k)$
  - $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n)$
  - $\mathbf{T}$ - 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 zaraďovacie pravidlo najväčšiu hodnotu

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

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

- Zaraďovacie 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 zaraďovacie pravidlo, ktoré zaradí 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 zaraďovania podľa maximálnej aposteriórnej pravdepodobnosti zaradíme 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 strátovu 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 $\mathbf{S} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{S}_i$

Analýza a klasifikace dat
Klasifikácia podľa vzdialeností
Miery vzdialeností a podobností

Institut biostatistiky a analýz
Masarykova univerzita

17. listopadu 2012

Základné pojmy

- Etalon - reprezentat triedy, voči ktorému určujeme vzdialenosť nových prvkov
- Rozhodovacie pravidlo

$$\omega_r = d(\mathbf{x}) = ||\mathbf{x}_{rE} - \mathbf{x}|| = \min_{\forall k} ||\mathbf{x}_{kE} - \mathbf{x}||$$

- T-prahová hodnota určujúca zaradenie do skupiny nezaraditeľných pozorovaní

$$d(\mathbf{x}) \leq T$$

- Vzdialenosť - miera nepodobnosti
- Podobnosť - duálna miera k vzdialenosti, čím sú si dva objekty bližšie, tým sú si podobnejšie

Metrika

Metrika je zobrazenie

$$\rho : \Theta \times \Theta \to R$$

- Θ - n-rozmerný obrazový priestor
- Predpoklady

$$\exists \rho_0 \in R : -\infty < \rho_0 \leq \rho(\mathbf{x}, \mathbf{y}) < \infty, \qquad \forall \mathbf{x}, \mathbf{y} \in \Theta$$

$$\rho(\mathbf{x}, \mathbf{y}) = 0, \quad \forall \mathbf{x} \in \Theta$$

- vlastnosti

$$\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in \Theta$$

$$\rho(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$$

- extra vlastnosť pre pravú metriku - trojuholníková nerovnosť

$$\rho(\mathbf{x}, \mathbf{z}) \leq \rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z}), \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \Theta$$

Metriky pre kvantitatívne znaky

- Euklidova metrika

$$\rho_E(\mathbf{x}_1, \mathbf{x}_2) = [\sum_{i=1}^{n}(x_{1i} - x_{2i})^2]^{\frac{1}{2}}$$

  - geometrický popis kružnice, zdôrazňuje väčšie rozdiely
  - bez odmocniny nesplňuje trojuholníkovú nerovnosť

- Hammingova metrika

$$\rho_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^{n}|x_{1i} - x_{2i}|$$

  - linearizácia euklidovskej, znížený význam členov s väčším rozdielom, znížená výpočetná náročnosť
  - dolný odhad euklidovskej vzdialenosti

Metriky pre kvantitatívne znaky

- Minkovského metrika

$$\rho_M(\mathbf{x}_1, \mathbf{x}_2) = [\sum_{i=1}^{n} |x_{1i} - x_{2i}|^m]^{\frac{1}{m}}$$

  - zobecnenie predchádzajúcich, zvyšuje váhu členov s väčším rozdielom
- Čebyševova

$$\rho_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} |x_{1i} - x_{2i}|$$

  - používa sa vo výpočetne náročných prípadoch
  - horný odhad Euklidovej vzdialenosti

Metriky pre kvantitatívne znaky

Nevýhody

- fyzikálne nezmyselne vytvárané súčty znakov rôznych jednotiek
- bez váhovania veľký vplyv korelovaných veličín
- Riešenie:
  - transformácia premenných
  - zavedenie vyrovnávacieho faktoru pre rôzne fyzikálne veličiny

Mahalanobisova metrika

- Metrika daná vzťahom

$$\rho_{MA}(\mathbf{x}_1, \mathbf{x}_2) = [(\mathbf{x}_1 - \mathbf{x}_2)' K^{-1} (\mathbf{x}_1 - \mathbf{x}_2)]^{\frac{1}{2}}$$

- váhové koeficienty sú dané inverznou kovariačnou maticou K^{-1}
- môže byť použitá tiež korelačná matica, kedy sa odstráni vplyv strednej hodnoty, teda sa odhalí informácia obsiahnutá vo variabilite dát

Metriky podobnosti

- Skalárny súčin v euklidovskom priestore

$$\sigma_s(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1'.\mathbf{x}_2 = \sum_{i=1}^{n} x_{1i} x_{2i}$$

  - miera podobnosti pre dva vektory s rovnakou dĺžkou
  - hodnota závisí na uhle a dĺžke vektorov
- Metrika kosínovej podobnosti - predpokladá normované vektory

$$\rho_{cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1' \mathbf{x}_2}{||\mathbf{x}_1||.||\mathbf{x}_2||}$$

- Pearsonov korelačný koeficient, Tanimotova metrika

Metriky vzdialeností pre kvalitatívne znaky

- Kontingenčá tabuľka pre dva vektory $\mathbf{x}, \mathbf{y}$ môže byť zapísaná ako matica A, v ktorej súradice vyjadrujú výskyt danej kombinácie znakov
- Hammingova metrika je definovaná počtom pozíc, v ktorých sa oba vektory líšia

$$\rho_{HQ}(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{k-1} \sum_{j=0 j,\neq i}^{k-1} a_{ij}$$

- pre binárne premenné

$$\rho_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i + y_i - 2x_i y_i)$$

Metriky podobnosti pre kvalitatívne znaky

- Majme množiny $X, Y, X \cap Y$, ich kardinalita nech je $n_X, n_Y, n_{X\cap Y}$, Tanimotova podobnosť množín X a Y je potom daná
$$\sigma_T(X, Y) = \frac{n_{X\cap Y}}{n_x + n_y - n_{X\cap Y}}$$
- pre dva vektory $\mathbf{x}, \mathbf{y}$ potom dostaneme
$$\sigma_{TQ}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_X + n_Y - n_{X\cap Y}}$$
- Pre dichotomické premenné: Jaccardov-Tanimotov asociačný koeficient, Dicov koeficient, Hamanov koeficient

Metriky vzdialeností medzi dvomi množinami

- Metóda najbližšieho suseda pre dva obrazy C_i, C_j

$$\rho_{NN}(C_i, C_j) = \min_{\mathbf{x}_p \in C_i, \mathbf{x}_q \in C_j} \rho(\mathbf{x}_p, \mathbf{x}_q)$$

- Metóda k najbližších susedov

$$\rho_{NNk}(C_i, C_j) = \min_{\mathbf{x}_p \in C_i, \mathbf{x}_q \in C_j} \sum^{k} \rho(\mathbf{x}_p, \mathbf{x}_q)$$

- Metóda najvzdialenejších susedov

$$\rho_{FN}(C_i, C_j) = \max_{\mathbf{x}_p \in C_i, \mathbf{x}_q \in C_j} \rho(\mathbf{x}_p, \mathbf{x}_q)$$

Metriky vzdialeností medzi dvomi množinami

- Centroidná metóda - vzdialenosť medzi množinami je daná vzdialenosťou ich reprezentatívnych obrazov (centroid, medoid)

$$\rho_{CE}(C_i, C_j) = \rho(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_j) = \sqrt{\sum_{s=1}^{n} (\bar{x}_{is} - \bar{x}_{js})^2}$$