Posplošena diskriminantna analiza z uporabo posplošenega singularnega razcepa

Jernej Banevec

Mentor: izred. prof. dr. Marjeta Knez

Fakulteta za matematiko in fiziko

dolga predstavitev diplomskega dela

3.4.2017

Pregled vsebine

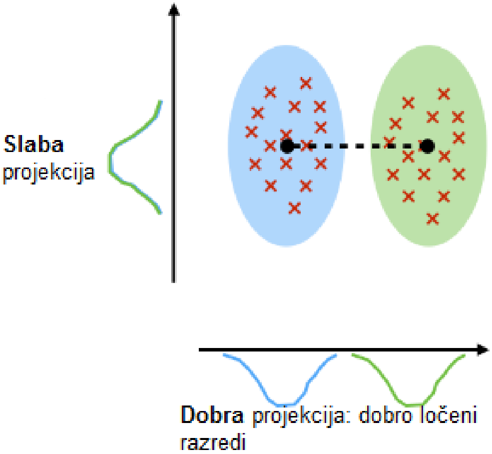
Posplošena diskriminantna analiza

- Posplošitev linearne diskriminantne analize
- Ena zelo uporabljenih statističnih metod
- Oblike podatkov:
    - Združeni v matriki $A \in \mathbb{R}^{mxn}$
    - m ...dimenzija posamezne meritve
    - n ...število meritev oz. podatkov
    - Podatki grupirani v k razredov oz. gruč

Posplošena diskriminantna analiza

- Iščemo preslikavo
$$G : \mathbb{R}^m \to \mathbb{R}^\ell,$$
kjer je $\ell \leqslant m - 1$
- Cilj:
    - Ohraniti razporejenost razredov
    - Zmanjšati razpršenost podatkov znotraj razredov
    - Povečati razlike med razredi

Primerjava dobra proti slabi preslikavi



Definicije

- Centroid i-tega razreda: $c^{(i)} = \frac{1}{n_i} \sum_{j \in N_i} a_j$
- Centroid celotnih podatkov: $c = \frac{1}{n} \sum_{j=1}^{n} a_j$
- $H_W = [A_1 - c^{(1)} e^{(1)^T}, \ldots, A_k - c^{(k)} e^{(k)^T}]$
- $H_B = [(c^{(1)} - c) e^{(1)^T}, \ldots, (c^{(k)} - c) e^{(k)^T}]$
- $e^{(i)} = (1, \ldots, 1)^T \in \mathbb{R}^{n_i x 1}$

Definicije

- Matrika razpršenosti podatkov znotraj razreda:

$$S_W = \sum_{i=1}^{k} \sum_{j \in N_i} (a_j - c^{(i)})(a_j - c^{(i)})^T = H_W H_W^T$$

- Matrika razlik med razredi:

$$S_B = \sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T = H_B H_B^T$$

- Matrika celotne razpršenosti:

$$S_M = S_W + S_B$$

- Tu so vse matrike elementi $\mathbb{R}^{mxm}$

Definicije

- Preslikava S_W na prostor dimenzije ℓ :

$$S_W^\ell = G^T S_W G$$

- Preslikava S_B na prostor dimenzije ℓ :

$$S_B^\ell = G^T S_B G$$

- Preslikava S_M na prostor dimenzije ℓ :

$$S_M^\ell = G^T S_M G$$

- Tu so vse matrike elementi $\mathbb{R}^{\ell x \ell}$
- Preslikamo tudi matriko podatkov A: $G^T A$

Posplošena diskriminantna analiza kot optimizacijski problem

- $sled(S_W) = \sum_{i=1}^{k} \sum_{j \in N_i} ||a_j - c^{(i)}||_2^2$
- $sled(S_B) = \sum_{i=1}^{k} n_i ||c^{(i)} - c||_2^2$
- Želimo:
  - povečati $sled(S_B^\ell)$
  - zmanjšati $sled(S_W^\ell)$
- Dobimo optimizacijski problem, pri katerem iščemo takšno preslikavo G, ki maksimizira

$$sled(G^T S_B G)/sled(G^T S_W G) \approx sled((S_W^\ell)^{-1} S_B^\ell)$$

- Kriterija ne moremo uporabiti ko S_W singularna oz. neobrnljiva

S_W obrnljiva

- Definicija:
  - $S_1 = S_B$
  - $S_2 = S_W$
- Tudi simetrično pozitivno definitna $\rightarrow$razcep Choleskega $S_2 = VV^T$
- Posplošen problem lastnih vrednosti $S_1 x = \lambda S_2 x$
- Označimo $J_1(G) = sled((G^T S_2 G)^{-1} G^T S_1 G)$
- Iščemo $G \in \mathbb{R}^{m x \ell}$, kjer $J_1(G)$ maksimalen
- $max_G J_1(G) \leq \lambda_1 + \lambda_2 + \ldots + \lambda_q = sled(S_2^{-1} S_1)$
- Optimalen $G = X \begin{bmatrix} I_\ell \\ 0 \end{bmatrix}$

Singularni razcep

- Originalna definicija posplošenega singularnega razcepa (Van Loan)

Izrek (Singularni razcep)

Za matriki $K_A \in \mathbb{R}^{pxm}$ z $p \geq m$ in $K_B \in \mathbb{R}^{nxm}$ obstajata ortogonalni matriki $U \in \mathbb{R}^{pxp}$ in $V \in \mathbb{R}^{nxn}$ ter nesingularna matrika $X \in \mathbb{R}^{mxm}$, da velja

$$U^T K_A X = diag(\alpha_1, ..., \alpha_m) \text{ in } V^T K_B X = diag(\beta_1, ..., \beta_q)$$

kjer $q = min(n, m)$, $\alpha_i \geq 0$ za $1 \leq i \leq m$ in $\beta_i \geq 0$ za $1 \leq i \leq q$.

Posplošeni singularni razcep

Izrek (Posplošeni singularni razcep)

Naj bosta dani matriki $K_A \in \mathbb{R}^{pxm}$ *in* $K_B \in \mathbb{R}^{nxm}$. *Potem za* $K = \begin{pmatrix} K_A \\ K_B \end{pmatrix}$ *in* $t = rang(K)$ *obstajajo ortogonalne matrike* $U \in \mathbb{R}^{pxp}$, $V \in \mathbb{R}^{nxn}$, $W \in \mathbb{R}^{txt}$ *in* $Q \in \mathbb{R}^{mxm}$, *da velja:*

$$U^T K_A Q = \Sigma_A \begin{pmatrix} W^T R, & 0 \end{pmatrix} \text{ in } V^T K_B Q = \Sigma_B \begin{pmatrix} W^T R, & 0 \end{pmatrix},$$

kjer je $\Sigma_A = \begin{pmatrix} I_A & & \\ & D_A & \\ & & 0_A \end{pmatrix}$ *in* $\Sigma_B = \begin{pmatrix} 0_B & & \\ & D_B & \\ & & I_B \end{pmatrix}$ *in*

Posplošeni singularni razcep nadaljevanje

Izrek (Posplošeni singularni razcep)

$R \in \mathbb{R}^{txt}$ nesingularna, matriki $I_A \in \mathbb{R}^{rxr}$ in $I_B \in \mathbb{R}^{(t-r-s)x(t-r-s)}$ identični matriki, kjer je

$$r = rang(K) - rang(K_B) \text{ in } s = rang(K_A) + rang(K_B) - rang(K),$$

$0_A \in \mathbb{R}^{(p-r-s)x(t-r-s)}$ in $0_B \in \mathbb{R}^{(n-t+r)xr}$,
$D_A = diag(\alpha_{r+1}, ..., \alpha_{r+s})$ in $D_B = diag(\beta_{r+1}, ..., \beta_{r+s})$, ki zadoščajo pogoju:

$$1 > \alpha_{r+1} \geq \ldots \geq \alpha_{r+s} > 0 \quad in \quad 1 < \beta_{r+1} \leq \ldots \leq \beta_{r+s} < 0$$

pri $\alpha_i^2 + \beta_i^2 = 1$ za $i = r+1, \ldots, r+s$

Posplošeni singularni razcep za nadaljno uporabo

Zgornja izreka povežemo s posplošitvijo singularnega razcepa Van Loana v obliko

$$U^T K_A X = \begin{pmatrix} \Sigma_A, & 0 \end{pmatrix} \text{in} V^T K_B X = \begin{pmatrix} \Sigma_B, & 0 \end{pmatrix},$$

kjer je

$$X = Q \begin{pmatrix} R^{-1} W & 0 \\ 0 & I \end{pmatrix}.$$

S_W poljubna

- G poiščemo z uporabo GSVD na $K = \begin{pmatrix} H_B^T \\ H_W^T \end{pmatrix} \in \mathbb{R}^{2nxm}$
- Dobimo
  - $H_B^T = U \begin{pmatrix} \Sigma_A, & 0 \end{pmatrix} X^{-1}$
  - $H_W^T = V \begin{pmatrix} \Sigma_B, & 0 \end{pmatrix} X^{-1}$
- Dodatno definirajmo:
  - $\alpha_i = 1, \beta_i = 0$ za $i = 1, 2, \ldots, r$
  - $\alpha_i = 0, \beta_i = 1$ za $i = r + s + 1, \ldots, t$
- Dobimo: $\beta_i^2 H_W H_W^T x_i = \alpha_i^2 H_B H_B^T x_i$

S_W poljubna

- H_W^T polnega ranga:
  - Rang torej m
  - S_W torej obrnljiva
  - $r = 0$ in $t = m$
  - $\beta_1, \ldots, \beta_m > 0$
  - Posplošeni problem lastnih vrednosti $H_W H_W^T x_i = \frac{\alpha_i^2}{\beta_i^2} H_B H_B^T x_i$
- H_W^T ni polnega ranga:
  - G tudi tu iz prvih ℓ stolpcev matrike X

Algoritem LDA/GSVD

Algoritem, ki za matriko podatkov $A \in \mathbb{R}^{mxn}$ s k razredi ustvari matriko $G \in \mathbb{R}^{mx\ell}$, ki ohranja oblike razredov v manj-dimenzionalnem prostoru in kjer $\ell < k$, z uporabo optimizacije

$$J_1(G) = sled((G^T S_W G)^{-1} G^T S_B G)$$

Algoritem LDA/GSVD

1. Iz A izračunamo H_B in H_W
2. Za $K = \begin{pmatrix} H_B^T \\ H_W^T \end{pmatrix} \in \mathbb{R}^{2nxm}$ izračunamo singularni razcep:

$$P^T KQ = \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix}$$

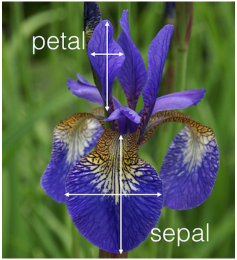
3. $t = rang(K)$
4. Izračunamo W iz singularnega razcepa za $P(1:n, 1:t)$:

$$U^T P(1:n, 1:t)W = \Sigma_A$$

5. G sestavimo iz prvih ℓ stolpcev matrike $X = \begin{pmatrix} R^{-1}W & 0 \\ 0 & I \end{pmatrix}$

Zgled - roža Iris

- Precej poznan zgled
- Trije razredi:
  1. Iris - setosa
  2. Iris - versicolor
  3. Iris - virginica



Zgled - roža Iris

- Podatki:

$$A = \begin{bmatrix} a_{1_{\text{sepal length}}} & a_{2_{\text{sepal length}}} & \cdots & a_{150_{\text{sepal length}}} \\ a_{1_{\text{sepal width}}} & a_{2_{\text{sepal width}}} & \cdots & a_{150_{\text{sepal width}}} \\ a_{1_{\text{petal length}}} & a_{2_{\text{petal length}}} & \cdots & a_{150_{\text{petal length}}} \\ a_{1_{\text{petal width}}} & a_{2_{\text{petal width}}} & \cdots & a_{150_{\text{petal width}}} \end{bmatrix}$$

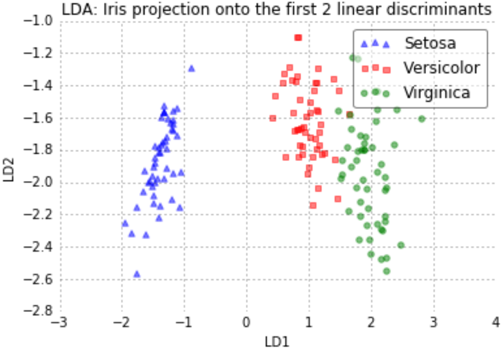
Zgled - roža Iris

- Na teh podatkih uporabimo posplošeno diskriminantno analizo

1. Računanje centroidov
2. Računanje matrik razpršenosti podatkov
3. Lastni razcep $(S_W)^{-1} S_B$
4. Urejanje lastnih vrednosti (po parih z vektorji)
5. Transformiranje na nov podprostor

Zgled - roža Iris

- Slikamo na 2-dimenzionalen (pod)prostor
- Dobimo sledečo sliko:



Viri

- Howland P. in Park H., Generalizing discriminant analysis using the generalizen singular value decomposition. TPAMI. 2004, 26, 8, 995-1006
- Jieping Ye, Characterization of a Family of Algorithms for Generalized Discriminant Analysis on Undersampled Problems. Journal of Machine Learning Research. 2005, 6, 483–502