Posplošena diskriminantna analiza z uporabo posplošenega singularnega razcepa

Jernej Banevec

Mentor: izred. prof. dr. Marjeta Knez

Fakulteta za matematiko in fiziko

kratka predstavitev diplomskega dela

8.11.2017

Pregled vsebine

Posplošena diskriminantna analiza

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

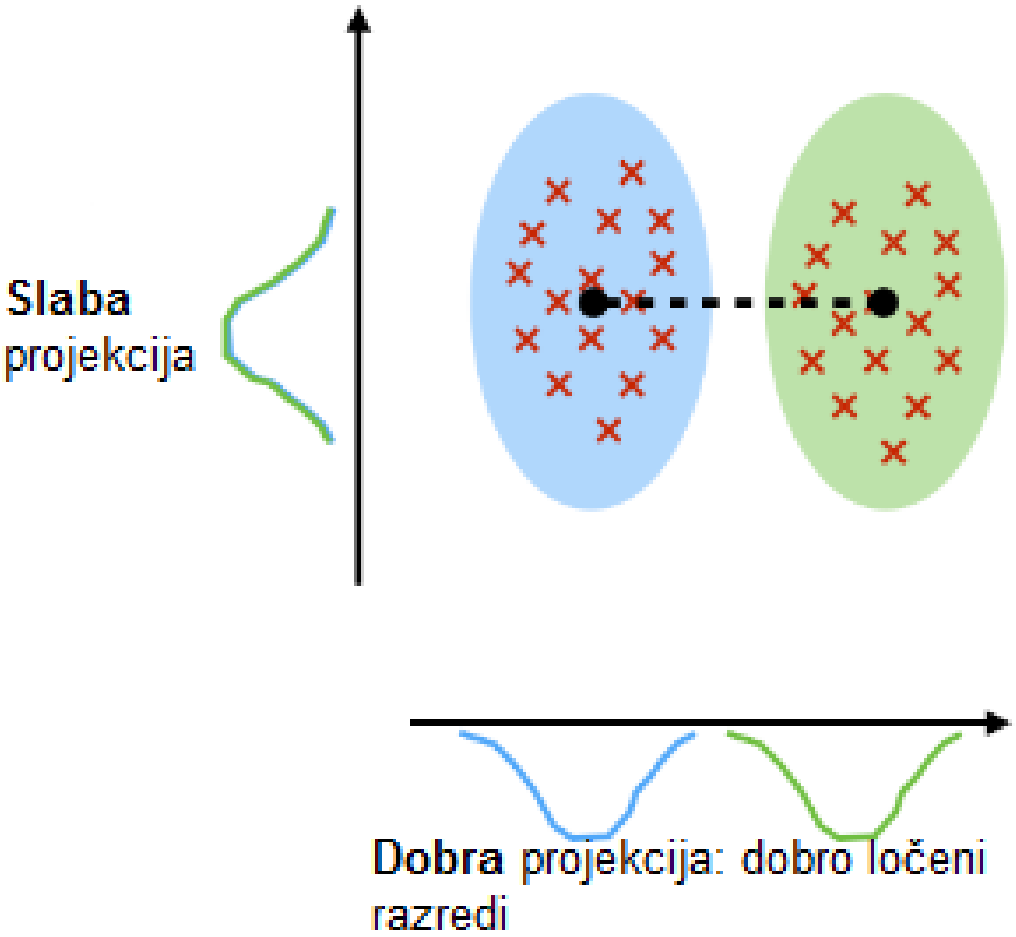
Posplošena diskriminantna analiza

- 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: $c = \frac{1}{n} \sum_{j=1}^{n} a_j$
- Matrika razpršenosti podatkov znotraj razreda:

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

- Matrika razlik med razredi:

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

- Matrika celotne razpršenosti:

$$S_M = S_W + S_B$$

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

Definicije

- Preslikava S_W na prostor dimenzije ℓ :

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

- Preslikava S_B na prostor dimenzije ℓ :

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

- Preslikava S_M na prostor dimenzije ℓ :

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

- Tu so vse matrike elementi $\mathbb{R}^{\ell x \ell}$
- $A^\ell = GA$

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

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

- Uporabno le ko je S_W^ℓ nesingularna oz. obrnljiva

Posplošeni singularni razcep

- Originalna definicija posplošenega singularnega razcepa (Van Loan)

Izrek (Posplošeni singularni razcep)

Za matriki $K_A \in \mathbb{R}^{p \times m}$ *z* $p \geqslant m$ *in* $K_B \in \mathbb{R}^{n \times m}$ *obstajata ortogonalni matriki* $U \in \mathbb{R}^{p \times p}$ *in* $V \in \mathbb{R}^{n \times n}$ *ter nesingularna matrika* $X \in \mathbb{R}^{m \times m}$, *da velja*

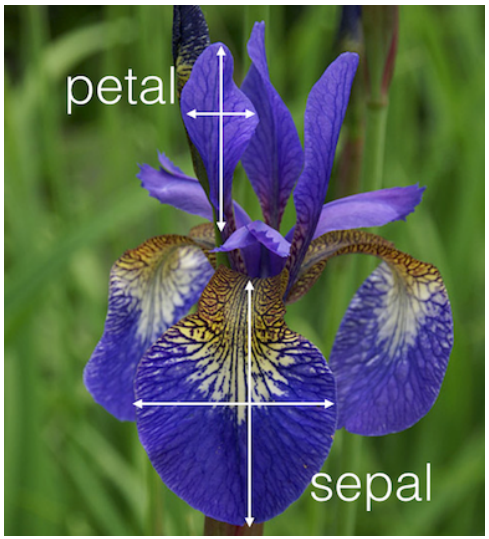
$$U^T K_A X = diag(\alpha_1, ..., \alpha_m) \text{ in } V^T K_B X = diag(\beta_1, ..., \beta_q)$$

kjer $q = min(n, m)$, $\alpha_i \geqslant 0$ *za* $1 \leqslant i \leqslant m$ *in* $\beta_i \geqslant 0$ *za* $1 \leqslant i \leqslant q$.

- Pozneje dodatno posplošimo singularni razcep

Zgled - roža Iris

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



Zgled - roža Iris

- Podatki:

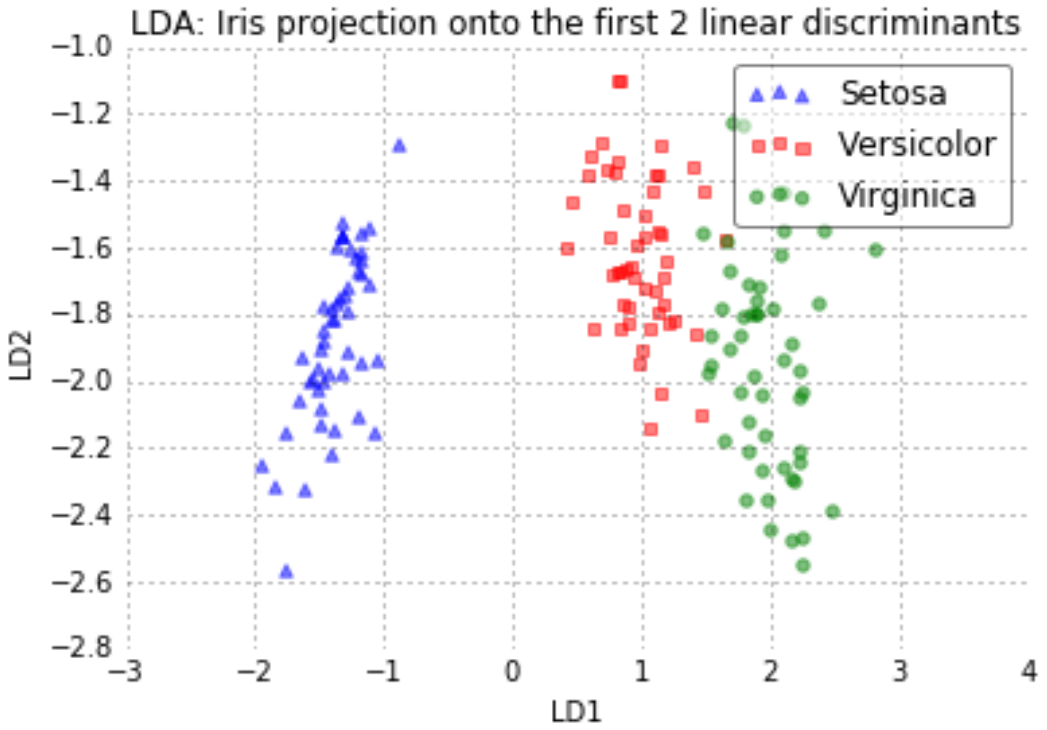
$$A = \begin{bmatrix} a1_{\text{sepal length}} & a2_{\text{sepal length}} & \cdots & a150_{\text{sepal length}} \\ a1_{\text{sepal width}} & a2_{\text{sepal width}} & \cdots & a150_{\text{sepal width}} \\ a1_{\text{petal length}} & a2_{\text{petal length}} & \cdots & a150_{\text{petal length}} \\ a1_{\text{petal width}} & a2_{\text{petal width}} & \cdots & a150_{\text{petal width}} \end{bmatrix}, Y = \begin{bmatrix} \omega_{\text{sitosa}} \\ \omega_{\text{sitosa}} \\ \vdots \\ \omega_{\text{versicolor}} \\ \vdots \\ \omega_{\text{virginica}} \end{bmatrix}$$

- Na teh podatkih uporabimo posplošeno diskriminantno analizo

1. Računanje centroidov
2. Računanje matrik razpršenosti podatkov
3. Lastni razcep $(S_W^\ell)^{-1} S_B^\ell$
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