

Optimalizace

8. PCA a úloha na nejmenší stopu

Tomáš Kroupa Tomáš Werner

2022 LS

Fakulta elektrotechnická
ČVUT v Praze

Databáze iris obsahuje $n = 150$ kosatců. U každého je uveden jeho druh a $m = 4$ charakteristiky jeho kališního/okvětního lístku.

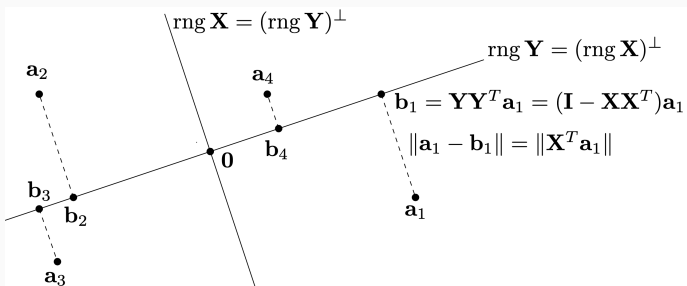
Redukce dimenze a vizualizace

Můžeme se snažit natrénovat klasifikátor kosatců, ale:

- Před tím je vhodné získat představu o povaze dat
- Datové vektory $\mathbf{a}_1, \dots, \mathbf{a}_{150} \in \mathbb{R}^4$ promítneme na vhodný podprostor dimenze k
- Souřadnice promítnutých bodů lze pro $k \leq 3$ zobrazit

Proložení bodů lineárním podprostorem

Pro vektory $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$ hledáme lineární podprostor $\text{rng } \mathbf{Y}$ dimenze $k \leq m$ minimalizující součet čtverců kolmých vzdáleností.



Úloha PCA pro $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_n]$

Minimalizuj $\sum_{i=1}^n \|\mathbf{X}^T \mathbf{a}_i\|^2$ za podmínky $\mathbf{X} \in \mathbb{R}^{m \times (m-k)}$, $\mathbf{X}^T \mathbf{X} = \mathbf{I}$

Co když prokládáme afinním podprostorem?

Tvrzení

Afinní podprostor dimenze k , který minimalizuje součet čtverců vzdáleností k vektorům $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{R}^m$, obsahuje jejich **těžiště**

$$\bar{\mathbf{a}} = \frac{1}{n}(\mathbf{a}_1 + \dots + \mathbf{a}_n).$$

1. Vektory \mathbf{a}_i posuneme tak, aby měly těžiště v $\mathbf{0}$:

$$\mathbf{a}_1 - \bar{\mathbf{a}}, \dots, \mathbf{a}_n - \bar{\mathbf{a}}$$

2. Posunuté vektory proložíme lin. podprostorem Y dimenze k
3. Hledaný afinní podprostor je $Y + \bar{\mathbf{a}}$

Proložení bodů lineárním podprostorem dimenze $k = m - 1$

To by měla být snadnější úloha: hledáme jednotkový vektor $\mathbf{x} \in \mathbb{R}^m$

Úloha

Minimalizuj $\sum_{i=1}^n |\mathbf{x}^T \mathbf{a}_i|^2 = \mathbf{x}^T \mathbf{A} \mathbf{A}^T \mathbf{x}$ za podmínky $\mathbf{x} \in \mathbb{R}^m$, $\|\mathbf{x}\| = 1$

- Spočítáme spektrální rozklad $\mathbf{A} \mathbf{A}^T = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$
- Ukážeme, že řešením úlohy je jednotkový vlastní vektor \mathbf{v}_1 odpovídající nejmenšímu vlastnímu číslu λ_1 matice $\mathbf{A} \mathbf{A}^T$
- *Hledanou ortonormální bázi* tedy tvoří komplementární vlastní vektory $\mathbf{v}_2, \dots, \mathbf{v}_m$

Věta (Courant–Fischer)

Nechť $\mathbf{B} \in \mathbb{R}^{m \times m}$ je symetrická s vlastními čísly $\lambda_1 \leq \dots \leq \lambda_m$.
Potom platí

$$\lambda_1 = \min \{ \mathbf{x}^T \mathbf{B} \mathbf{x} \mid \mathbf{x} \in \mathbb{R}^m, \|\mathbf{x}\| = 1 \}$$

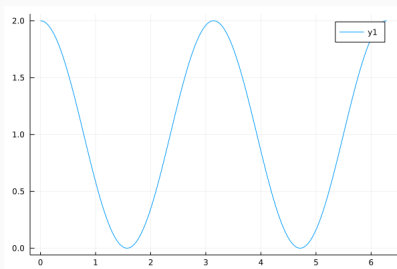
a minima se nabývá pro vlastní vektor \mathbf{v}_1 odpovídající λ_1 .

Zbývající vlastní čísla pro $k = 2, \dots, m$ lze popsat takto:

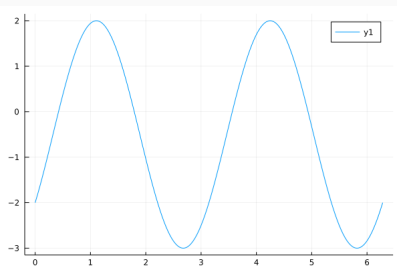
$$\lambda_k = \min \{ \mathbf{x}^T \mathbf{B} \mathbf{x} \mid \mathbf{x} \in \mathbb{R}^m, \|\mathbf{x}\| = 1, \mathbf{v}_1^T \mathbf{x} = \dots = \mathbf{v}_{k-1}^T \mathbf{x} = 0 \}$$

Příklady funkce $x^T Bx$ v polárních souřadnicích

- $B = \text{diag}(2, 0)$
- Vlastní čísla 0 a 2
- Vlastní vektory $(0, 1)$ a $(1, 0)$



- $B = \begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$
- Vlastní čísla -3 a 2
- Vl.vektory $(-2, 1)$ a $(1, 2)$



- Předchozí postup *nelze použít*, platí totiž jen

$$\sum_{i=1}^n \|\mathbf{X}^T \mathbf{a}_i\|^2 = \sum_{j=1}^{m-k} \mathbf{x}_j^T \mathbf{A} \mathbf{A}^T \mathbf{x}_j$$

kde \mathbf{x}_j jsou sloupce hledané matice \mathbf{X}

- Pomocí tzv. *stopy matice* můžeme vyjádřit

$$\sum_{j=1}^{m-k} \mathbf{x}_j^T \mathbf{A} \mathbf{A}^T \mathbf{x}_j = \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{A}^T \mathbf{X})$$

Reformulace problému tak povede na *úlohu o nejmenší stopě*.

Stopa

Stopa čtvercové matice $\mathbf{A} \in \mathbb{R}^{n \times n}$ je číslo

$$\operatorname{tr} \mathbf{A} = a_{11} + \cdots + a_{nn}.$$

Vlastnosti

1. $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr} \mathbf{A} + \operatorname{tr} \mathbf{B}$, $\operatorname{tr}(\alpha \mathbf{A}) = \alpha \operatorname{tr} \mathbf{A}$
2. $\operatorname{tr}(\mathbf{A}^T) = \operatorname{tr} \mathbf{A}$
3. $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$, kde $\mathbf{A} \in \mathbb{R}^{m \times n}$ a $\mathbf{B} \in \mathbb{R}^{n \times m}$
4. $\operatorname{tr} \mathbf{A} = \lambda_1 + \cdots + \lambda_n$

Stopy některých matic

Symetrická pozitivně definitní matice $\mathbf{A} \in \mathbb{R}^{2 \times 2}$

Má vlastní čísla $\lambda_1, \lambda_2 > 0$ a jednotkové vlastní vektory $\mathbf{v}_1 \perp \mathbf{v}_2$.
Obvod obdélníku o stranách $\mathbf{A}\mathbf{v}_1$ a $\mathbf{A}\mathbf{v}_2$ je

$$2(\lambda_1 + \lambda_2) = 2 \operatorname{tr}(\mathbf{A}).$$

Ortogonální projektor

Nechť $\mathbf{U} \in \mathbb{R}^{m \times k}$ je matice s ortonormálními sloupci. Ortogonální projektor $\mathbf{P} = \mathbf{U}\mathbf{U}^T$ na podprostor $\operatorname{rng} \mathbf{U}$ dimenze k má stopu

$$\operatorname{tr} \mathbf{P} = k.$$

Skalární součin matic $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ je číslo

$$\langle \mathbf{A}, \mathbf{B} \rangle = \text{tr}(\mathbf{A}^T \mathbf{B}).$$

Vlastnosti

$$\begin{aligned} \langle \mathbf{A}, \mathbf{B} \rangle &= \sum_{i=1}^m \sum_{j=1}^n a_{ij} b_{ij} \\ &= \text{tr}(\mathbf{B}^T \mathbf{A}) = \text{tr}(\mathbf{A} \mathbf{B}^T) = \text{tr}(\mathbf{B} \mathbf{A}^T) \end{aligned}$$

Norma matice

Frobeniova norma matice $\mathbf{A} \in \mathbb{R}^{m \times n}$ je

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}.$$

Vlastnosti

$$\|\mathbf{A}\| = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^T)} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} = \sqrt{\lambda_1 + \dots + \lambda_m},$$

kde $\lambda_i \geq 0$ jsou vlastní čísla pozitivně semidefinitní matice $\mathbf{A}\mathbf{A}^T$.

Vzdálenost matic $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ definujeme jako $\|\mathbf{A} - \mathbf{B}\|$.

PCA jako úloha na nejmenší stopu

Formulace PCA pomocí stopy

Úloha PCA pro $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_n]$

Minimalizuj $\sum_{j=1}^{m-k} \mathbf{x}_j^T \mathbf{A} \mathbf{A}^T \mathbf{x}_j$ za podmínky $\mathbf{X} \in \mathbb{R}^{m \times (m-k)}$, $\mathbf{X}^T \mathbf{X} = \mathbf{I}$

Definice stopy dává

$$\sum_{j=1}^{m-k} \mathbf{x}_j^T \mathbf{A} \mathbf{A}^T \mathbf{x}_j = \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{A}^T \mathbf{X})$$

Ekvivalentní formulace úlohy PCA

$$\min \left\{ \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{A}^T \mathbf{X}) \mid \mathbf{X} \in \mathbb{R}^{m \times (m-k)}, \mathbf{X}^T \mathbf{X} = \mathbf{I} \right\}$$

Úloha na nejmenší stopu

Věta

Nechť $\mathbf{B} \in \mathbb{R}^{m \times m}$ je symetrická matice se spektrálním rozkladem $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, vlastními čísly $\lambda_1 \leq \dots \leq \lambda_m$ a $\ell \leq m$. Platí

$$\min \left\{ \text{tr}(\mathbf{X}^T \mathbf{B} \mathbf{X}) \mid \mathbf{X} \in \mathbb{R}^{m \times \ell}, \mathbf{X}^T \mathbf{X} = \mathbf{I} \right\} = \lambda_1 + \dots + \lambda_\ell$$

a minima se nabývá pro $\mathbf{X} = [\mathbf{v}_1 \cdots \mathbf{v}_\ell]$.

Pro $\ell = 1$ je to Courant-Fischerova věta.

PCA jako úloha na nejmenší stopu

$$\min \left\{ \text{tr}(\mathbf{X}^T \mathbf{A} \mathbf{A}^T \mathbf{X}) \mid \mathbf{X} \in \mathbb{R}^{m \times (m-k)}, \mathbf{X}^T \mathbf{X} = \mathbf{I} \right\}$$

1. Pro matici $\mathbf{A} \mathbf{A}^T \in \mathbb{R}^{m \times m}$ spočítáme spektrální rozklad $\mathbf{V} \mathbf{\Lambda} \mathbf{V}^T$ a seřadíme vzestupně vlastní čísla, $\lambda_1 \leq \dots \leq \lambda_m$
2. Označíme $\mathbf{V} = [\underbrace{\mathbf{v}_1 \cdots \mathbf{v}_{m-k}}_{\mathbf{X}} \underbrace{\mathbf{v}_{m-k+1} \cdots \mathbf{v}_m}_{\mathbf{Y}}]$
3. Sloupce matice $\mathbf{Y} \in \mathbb{R}^{m \times k}$ jsou ortonormální bázi hledaného podprostoru dimenze k
4. Optimální hodnota úlohy $\lambda_1 + \dots + \lambda_{m-k}$ je *chyba proložení*

Příklad – iris (1)

Matice $\mathbf{A} \in \mathbb{R}^{4 \times 150}$ má v každém z $n = 150$ sloupců měření $m = 4$ proměnných, od nichž jsme odečetli $\bar{\mathbf{a}}$. Volíme dimenzi $k = 2$.

Řešení

- $\mathbf{AA}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, kde $\mathbf{\Lambda} = \text{diag}(3.53, 11.70, 36.10, 629.50)$
- Ortonormální báze hledaného podprostoru je

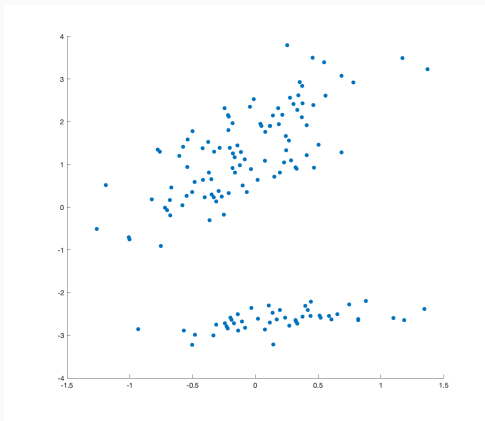
$$\mathbf{Y} = \begin{bmatrix} \mathbf{v}_3 & \mathbf{v}_4 \end{bmatrix} \in \mathbb{R}^{4 \times 2}$$

- Relativní chyba proložení je

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \approx 0.02$$

Příklad – iris (2)

- Chceme vizualizovat první dvě hlavní komponenty
- Zobrazíme si souřadnice promítnutých bodů v \mathbb{R}^2
- Nalezneme je ve sloupcích matice $\mathbf{Y}^T \mathbf{A} \in \mathbb{R}^{2 \times 150}$

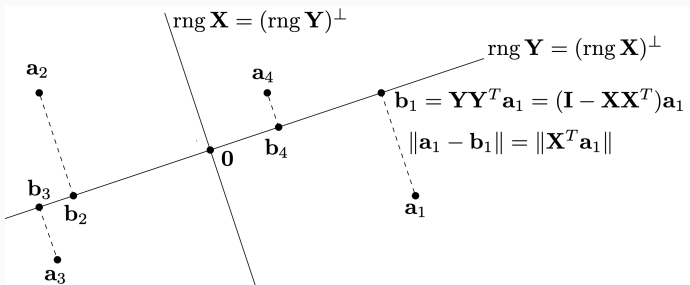


Nejbližší matice nižší hodnosti

Tato úloha je ekvivalentní úloze PCA:

Low rank approximation pro matici $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_n]$

$$\min \{ \|\mathbf{A} - \mathbf{B}\|^2 \mid \mathbf{B} \in \mathbb{R}^{m \times n}, \text{rank } \mathbf{B} \leq k \}$$



Optimální řešení je $\mathbf{B} = \mathbf{Y}\mathbf{Y}^T\mathbf{A}$.

- Ortonormální báze nalezeného podprostoru je v $\mathbf{Y} \in \mathbb{R}^{m \times k}$
- Ortogonální projekce \mathbf{a}_i na ten podprostor je $\mathbf{b}_i = \mathbf{Y}\mathbf{Y}^T\mathbf{a}_i$
- Vektor souřadnic bodu \mathbf{b}_i v ortonormální bázi \mathbf{Y} je $\mathbf{Y}^T\mathbf{a}_i$
- Matice souřadnic těch bodů je $\mathbf{Y}^T\mathbf{A} \in \mathbb{R}^{k \times n}$

Použití

1. **Kompresce:** \mathbf{A} má mn prvků, \mathbf{Y} a $\mathbf{Y}^T\mathbf{A}$ mají $(m+n)k$ prvků
2. **Redukce dimenze:** Body $\mathbf{Y}^T\mathbf{A}$ jsou v menší dimenzi než \mathbf{A}
3. **Vizualizace:** Pro $k \leq 3$ si lze body $\mathbf{Y}^T\mathbf{A}$ zobrazit
4. **Rozpoznávání:** Body $\mathbf{Y}^T\mathbf{A}$ jsou vhodnější pro klasifikaci atp.