Optimalizace

8. PCA a úloha na nejmenší stopu

Tomáš Kroupa Tomáš Werner

2023 LS

Fakulta elektrotechnická
ČVUT v Praze

Motivace

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

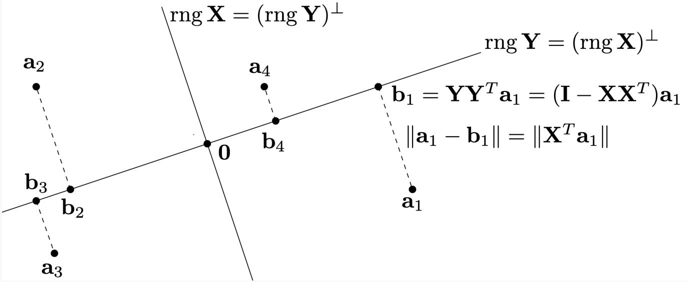
Redukce dimenze a vizualizace

Můžeme se snažit natrénovat klasifikátor kosatců, ale:

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

Proložení bodů lineárním podprostorem

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



Úloha PCA pro $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix} \in \mathbb{R}^{m \times n}$

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

Co když prokládáme afinním podprostorem?

Tvrzení

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

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

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

$$\mathbf{a}_1 - \overline{\mathbf{a}}, \ldots, \mathbf{a}_n - \overline{\mathbf{a}}$$

2. Posunuté vektory proložíme lineárním prostorem Y dimenze k
3. Hledaný afinní podprostor je $Y + \overline{\mathbf{a}}$

Úloha PCA pro $k = m - 1$

To by měla být snadnější úloha, protože hledaná matice $\mathbf{X} \in \mathbb{R}^{m \times (m-k)}$ je typu $m \times 1$, tedy vlastně jen vektor $\mathbf{x} \in \mathbb{R}^m$:

Úloha

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

- Hodí se vyjádřit $\sum_{i=1}^{n} (\mathbf{x}^\mathsf{T} \mathbf{a}_i)^2 = \mathbf{x}^\mathsf{T} \mathbf{A} \mathbf{A}^\mathsf{T} \mathbf{x}$
- Optimální řešení $\mathbf{x}^*$ úlohy je kolmé na hledaný podprostor dimenze $m - 1$

Optimalizační popis vlastních čísel

Věta (Courant–Fischer)

Nechť $\mathbf{B} \in \mathbb{R}^{m \times m}$ je symetrická s vlastními čísly $\lambda_1 \leq \cdots \leq \lambda_m$ a ortonormální bází vlastních vektorů $\mathbf{v}_1, \ldots, \mathbf{v}_m$. Potom platí

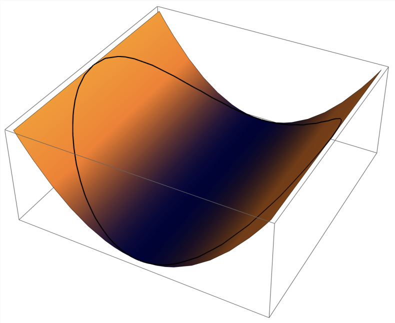
$$\min \{\mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^m,\ \|\mathbf{x}\| = 1\} = \lambda_1,$$
$$\max \{\mathbf{x}^\mathsf{T}\mathbf{B}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^m,\ \|\mathbf{x}\| = 1\} = \lambda_m,$$

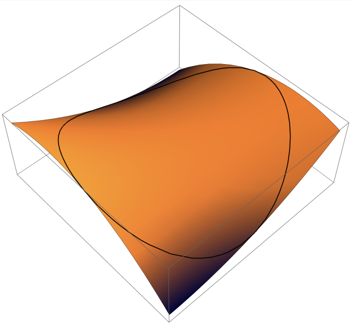
minima se nabývá pro $\mathbf{v}_1$ a maxima pro $\mathbf{v}_m$.

Příklady

- $\mathbf{B} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$
- Forma $2x_1^2$
- Vlastní čísla 0 a 2
- Vlastní vektory $(0,1)$ a $(1,0)$

- $\mathbf{B} = \begin{bmatrix} -2 & 2 \\ 2 & 1 \end{bmatrix}$
- Forma $-2x_1^2 + x_2^2 + 4x_1x_2$
- Vlastní čísla -3 a 2
- Vl.vektory $(-2,1)$ a $(1,2)$





Řešení úlohy PCA pro $k = m - 1$

Úloha

Minimalizuj $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{x}$ za podmínky $\mathbf{x} \in \mathbb{R}^m$, $\|\mathbf{x}\| = 1$

Řešení vyčteme ze spektrálního rozkladu $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$:

- Optimální řešení je $\mathbf{x}^* = \mathbf{v}_1$
- Tedy hledaný podprostor dimenze $m - 1$ má bázi $\mathbf{v}_2, \dots, \mathbf{v}_m$
- Chyba proložení je λ_1

Úloha PCA pro $k < m - 1$

- Předchozí postup *nelze přímo použít*, protože platí jen

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

- Ale můžeme vyjádřit

$$\sum_{i=1}^{n} \|\mathbf{X}^\mathsf{T}\mathbf{a}_i\|^2 = \operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{X})$$

a vyřešit PCA pomocí *úlohy o nejmenší stopě*

Stopa

Stopa matice

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

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

Vlastnosti

1. $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr} \mathbf{A} + \operatorname{tr} \mathbf{B}, \quad \operatorname{tr}(\alpha \mathbf{A}) = \alpha \operatorname{tr} \mathbf{A}$
2. $\operatorname{tr}(\mathbf{A}^\mathsf{T}) = \operatorname{tr} \mathbf{A}$
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. $\operatorname{tr} \mathbf{A} = \lambda_1 + \cdots + \lambda_n$

Skalární součin matic

Pro matice $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ definujeme skalární součin

$$\langle \mathbf{A}, \mathbf{B} \rangle := \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}.$$

Vlastnosti

- Pro $n = 1$ platí $\langle \mathbf{A}, \mathbf{B} \rangle = \mathbf{a}^{\mathsf{T}} \mathbf{b}$.
- Platí $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}^{\mathsf{T}} \mathbf{B})$.

Norma matice

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

$$\|\mathbf{A}\| := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}.$$

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

Vlastnosti

$$\|\mathbf{A}\| = \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle} = \sqrt{\operatorname{tr}(\mathbf{A}\mathbf{A}^\mathsf{T})} = \sqrt{\lambda_1 + \cdots + \lambda_m},$$

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

PCA jako úloha na nejmenší stopu

Formulace PCA pomocí stopy

Původní formulace PCA pro $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix} \in \mathbb{R}^{m\times n}$

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

Z definice stopy plyne

$$\sum_{i=1}^{n} \|\mathbf{X}^\mathsf{T}\mathbf{a}_i\|^2 = \operatorname{tr}(\mathbf{X}^\mathsf{T}\mathbf{A}\mathbf{A}^\mathsf{T}\mathbf{X}).$$

Ekvivalentní formulace úlohy PCA

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

Úloha na nejmenší stopu

Věta

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

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

a minima se nabývá pro $\mathbf{X} = \begin{bmatrix} \mathbf{v}_1 \cdots \mathbf{v}_\ell \end{bmatrix}$.

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

PCA kuchařka

PCA jako instance úlohy na nejmenší stopu

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

1. Spočti spektrální rozklad $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}$, kde použiješ řazení $\lambda_1 \leq \cdots \leq \lambda_m$ a $\mathbf{V} = [\underbrace{\mathbf{v}_1 \cdots \mathbf{v}_{m-k}}_{\mathbf{X}}\ \underbrace{\mathbf{v}_{m-k+1} \cdots \mathbf{v}_m}_{\mathbf{Y}}]$
2. Ve sloupcích matice $\mathbf{Y} \in \mathbb{R}^{m\times k}$ najdeš ortonormální bázi hledaného podprostoru dimenze k
3. Optimální hodnota úlohy $\lambda_1 + \cdots + \lambda_{m-k}$ je *chyba proložení*

Malý příklad

Vektory $\mathbf{a}_1 = (1,3,0)$, $\mathbf{a}_2 = (2,1,1)$, $\mathbf{a}_3 = (-1,3,0)$ a $\mathbf{a}_4 = (2,-3,0)$ se zřejmě příliš neliší v poslední souřadnici. Přesvědčí nás o tom PCA pro dimenzi $k = 2$.

- Vezmeme matici vystředěných vektorů $\mathbf{A}$
- $\mathbf{A}\mathbf{A}^\mathsf{T} = \begin{bmatrix} 6 & -8 & 1 \\ -8 & 24 & 0 \\ 1 & 0 & 0.75 \end{bmatrix} = \mathbf{V}\,\mathrm{diag}(0.4, 3.3, 27.0)\mathbf{V}^\mathsf{T}$
- Chyba proložení podprostorem s bází $\mathbf{v}_2, \mathbf{v}_3$ je $\lambda_1 = 0.4$
- Relativní chyba proložení $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} \approx 0.1$

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 $k = 2$.

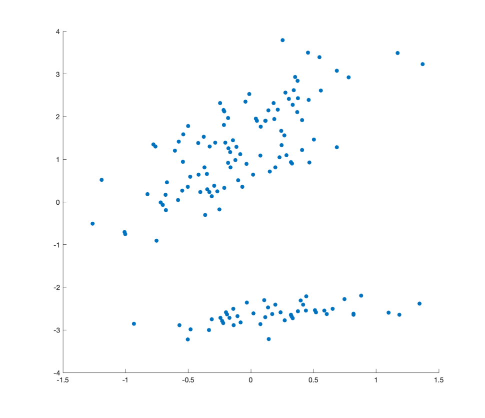
Řešení

- $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}$, kde $\boldsymbol{\Lambda} = \operatorname{diag}(3.53,\ 11.70,\ 36.10,\ 629.50)$
- Hledaný podprostor má bázi $\mathbf{Y} = \begin{bmatrix} \mathbf{v}_3 & \mathbf{v}_4 \end{bmatrix} \in \mathbb{R}^{4\times 2}$
- Chyba je $\lambda_1 + \lambda_2$
- 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 $\mathbf{Y}^\top \mathbf{A} \in \mathbb{R}^{2\times 150}$



Shrnutí PCA

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

Aplikace

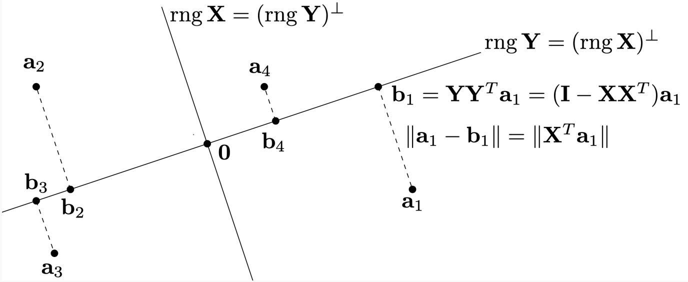
1. Komprese: $\mathbf{A}$ má mn prvků, $\mathbf{Y}$ a $\mathbf{Y}^\mathsf{T}\mathbf{A}$ mají $(m+n)k$ prvků
2. Redukce dimenze: Body $\mathbf{Y}^\mathsf{T}\mathbf{A}$ jsou v menší dimenzi než $\mathbf{A}$
3. Vizualizace: Pro $k \leq 3$ si lze body $\mathbf{Y}^\mathsf{T}\mathbf{A}$ zobrazit
4. Rozpoznávání: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

Nejbližší matice nižší hodnosti

Tato úloha je ekvivalentní úloze PCA:

Low rank approximation pro matici $\mathbf{A} = \begin{bmatrix}\mathbf{a}_1 \cdots \mathbf{a}_n\end{bmatrix}$

$$\min\left\{\|\mathbf{A} - \mathbf{B}\|^2 \mid \mathbf{B} \in \mathbb{R}^{m\times n},\ \operatorname{rank} \mathbf{B} \le k\right\}$$



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