Optimalizace

8. PCA a úloha na nejmenší stopu

Tomáš Kroupa Tomáš Werner

2021 LS

Fakulta elektrotechnická
ČVUT v Praze

Motivace

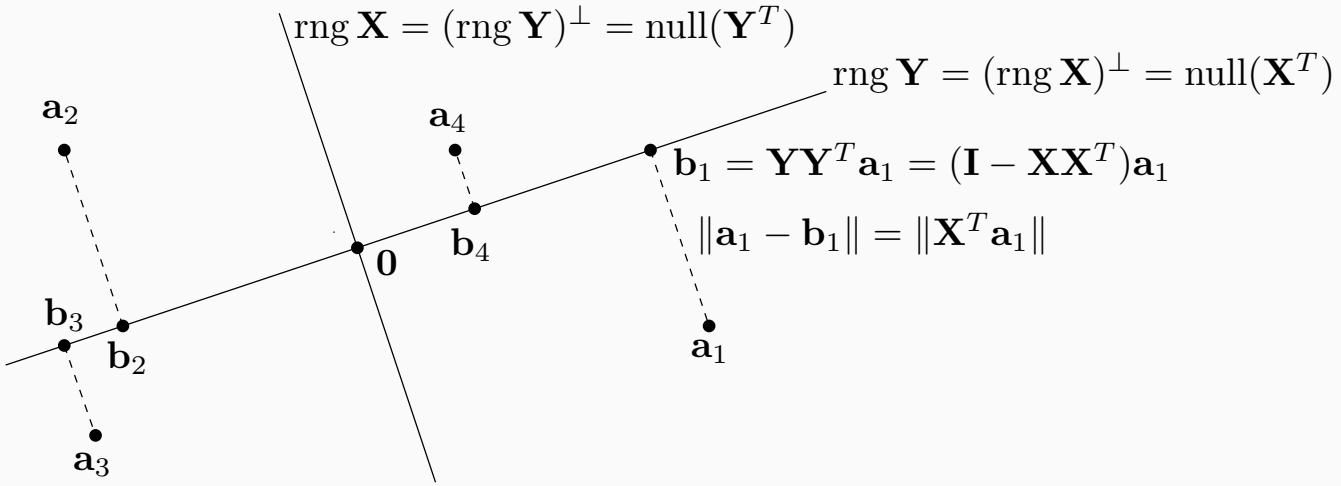
Databáze `iris` obsahuje $n = 150$ kosatců. U každého je uveden jeho druh a $m = 4$ charakteristiky jeho kališního/okvětního lístku.

Redukce dimenze a vizualizace

- Můžeme se snažit natrénovat klasifikátor kosatců, ale:
- Před tím je vhodné získat představu o povaze dat
- Naměřené vektory v $\mathbb{R}^4$ promítneme na podprostor dimenze k
- 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 $\mathbf{A} \coloneqq \begin{bmatrix} \mathbf{a}_1 \cdots \mathbf{a}_n \end{bmatrix}$

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

Co když prokládáme afinním podprostorem?

Tvrzení

Afinní podprostor dimenze k, který minimalizuje součet čtverců vzáleností 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. Zadané vektory posuneme tak, aby jejich těžiště byl $\mathbf{0}$:

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

2. Proložíme je lineárním podprostorem Y dimenze k
3. Hledaný afinní podprostor je $Y + \overline{\mathbf{a}}$

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

To by měla být snadnější úloha: hledaná matice je $\mathbf{X} \in \mathbb{R}^{m \times 1}$

Úloha

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

- Spektrální rozklad $\mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$
- Úlohu řeší jednotkový vlastní vektor $\mathbf{v}_1$ odpovídající nejmenšímu vlastnímu číslu λ_1 matice $\mathbf{A}\mathbf{A}^\mathsf{T}$ – viz další slajd
- *Hledanou ortonormální bázi* tvoří vlastní vektory $\mathbf{v}_2, \ldots, \mathbf{v}_m$

Optimalizační popis vlastních čísel

Věta (Courant–Fischer)

Nechť $\mathbf{B} \in \mathbb{R}^{m \times m}$ je symetrická matice a seřaďme její vlastní čísla vzestupně, $\lambda_1 \leq \cdots \leq \lambda_m$. Potom platí

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

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

Zobecněním úlohy lze získat vlastní čísla pro $2 \leq k \leq m$:

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

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

- Předchozí postup *nelze použít*, platí totiž jen

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

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

- To je součet hodnot kvadratické formy na ortonormální množině a pomocí *stopy matice* ho můžeme zapsat jako

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

Reformulace problému tak povede na *úlohu 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. $\operatorname{tr}(\mathbf{A} + \mathbf{B}) = \operatorname{tr} \mathbf{A} + \operatorname{tr} \mathbf{B}$, $\operatorname{tr}(\alpha \mathbf{A}) = \alpha \operatorname{tr} \mathbf{A}$
2. $\operatorname{tr}(\mathbf{A}^{\mathsf{T}}) = \operatorname{tr} \mathbf{A}$
3. $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$ pro každé $\mathbf{A} \in \mathbb{R}^{m \times n}$ 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 $\mathbf{A} \in \mathbb{R}^2$

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

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

Ortogonální projektor

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

$$\mathrm{tr}\,\mathbf{P} = k.$$

Skalární součin matic

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

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

Vlastnosti

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

Norma matice

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

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

Vlastnosti

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

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

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 $\mathbf{A} := \begin{bmatrix} \mathbf{a}_1 \cdots \mathbf{a}_n \end{bmatrix}$

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

Definice stopy dává

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

Ekvivalentní formulace úlohy PCA

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

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

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

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

PCA – řešení

PCA jako úloha 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. Pro matici $\mathbf{A}\mathbf{A}^\mathsf{T} \in \mathbb{R}^{m \times m}$ spočítáme spektrální rozklad $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$ a seřadíme vzestupně vlastní čísla, $\lambda_1 \leq \cdots \leq \lambda_m$
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 bází hledaného podprostoru dimenze k
4. Optimální hodnota úlohy $\lambda_1 + \cdots + \lambda_{m-k}$ 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

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

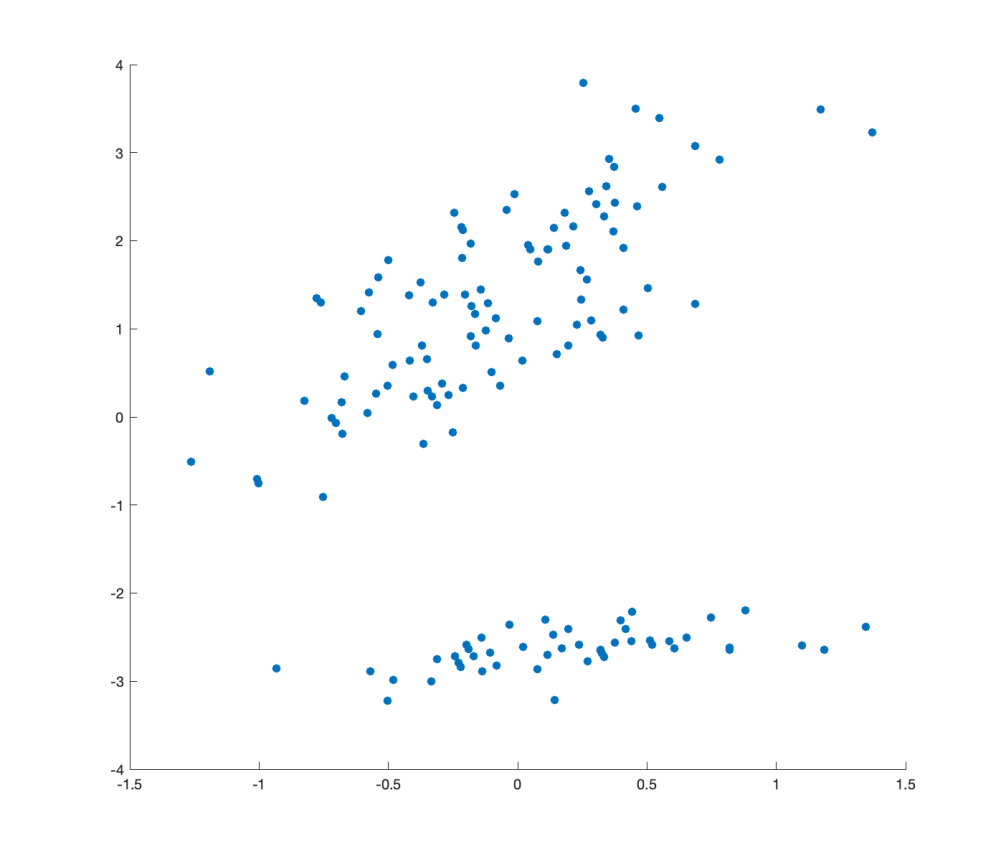
$$\mathbf{Y} = \begin{bmatrix} \mathbf{v}_3 & \mathbf{v}_4 \end{bmatrix} \in \mathbb{R}^{4\times 2}$$

- Chyba proložení je relativně malá, neboť

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{4} \lambda_i} \approx 0.02$$

Příklad – `iris` (2)

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

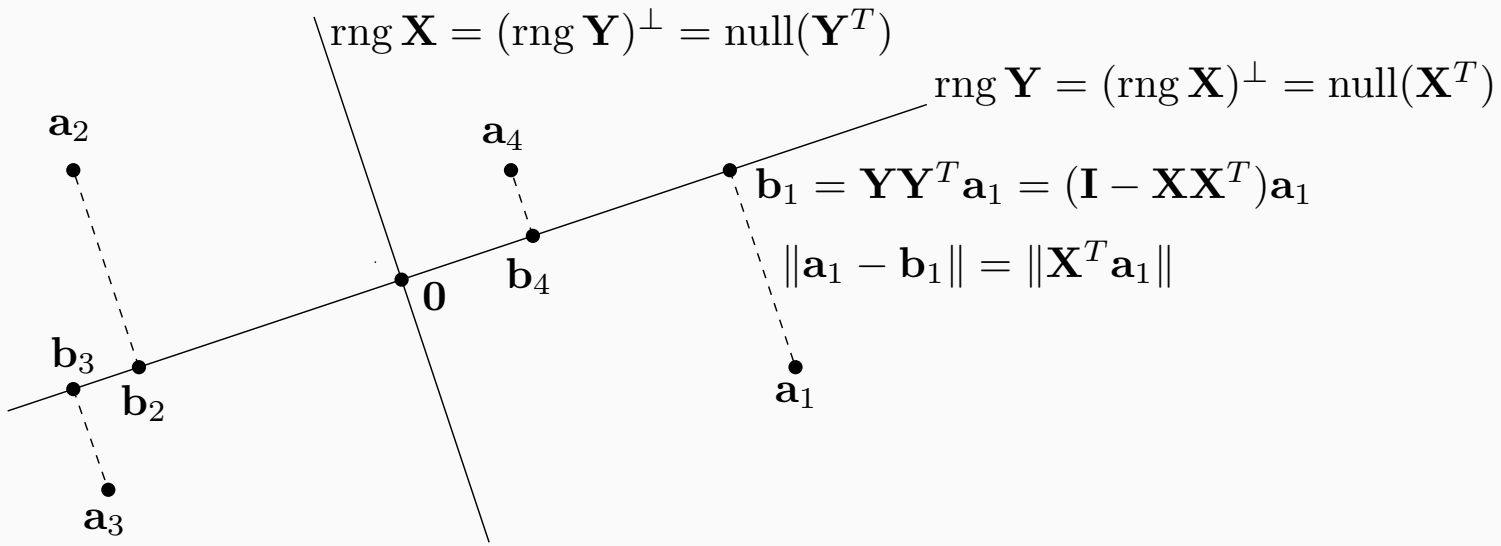


Nejbližší matice nižší hodnosti

Tato úloha je ekvivalentní úloze PCA:

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

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



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

Aplikace PCA

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

$$\mathbf{C} = [\mathbf{c}_1 \dots \mathbf{c}_n] = \mathbf{Y}^\mathsf{T}\mathbf{A} \in \mathbb{R}^{k \times n}$$

Použití

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

Reference


T. Werner. *Optimalizace* (kapitola 7). Elektronická skripta. FEL ČVUT, 2020.