Domácí úkol OPT 3

by a very good student

13. listopadu 2019

Volby do poslanecké sněmovny 2017

Občanské sdružení KohoVolit.eu připravilo pro poslední volby do Poslanecké sněmovny parlamentu ČR tzv. Volební kalkulačku. Základní myšlenka volební kalkulačky je pomoci voliči, aby se rychle zorientoval v programu politických stran, a to tak, že se měří shoda vlastních voličových názorů s programovými otázkami kandidátních stran prostřednictvím sady otázek z mnoha oblastí.

Byla vytvořena sada 100 otázek, na které lze odpovědět Ano (1)/Ne (-1)/Nevím (0). Kandidující politické strany byly obeslány, 25 z nich dodalo odpovědi. Volič na webu následně odpoví na stejné otázky a volební kalkulačka vyhodnotí shodu jeho odpovědí a seřadí a doporučí politické strany. Existuje i kratší verze kalkulačky, kde je jen podmnožina 35 otázek. My se ale nadále budeme zabývat úplnou sadou 100 otázek.

Úkol 1

Jednotlivé politické strany $\mathbf{a}_i \in \mathbb{R}^{100}$, $i \in \{1, ..., 25\}$ je možné chápat jako body v prostoru odpovědí na referenční programové otázky. Chtěli bychom tento prostor vhodně vizualizovat. To uděláme tak, že body a_i proložíme afinním podprostorem dimenze 2 tak, aby součet kvadrátů vzdáleností původních \mathbf{a}_i a promítnutých bodů \mathbf{a}_i' byl minimální. Následně promítnuté body zobrazíme v souřadnicích báze nalezeného afinního podprostoru.

Formulujte optimalizační problém.

Hledáme afinní podprostor dimenze 2 prostoru \mathbb{R}^{100} , který minimalizuje součet čtverců vzdáleností k daným bodům $a_1, ..., a_{25}$.

pozn. opravujícího: Chybí zde formální zápis min $||...||^2$ za podmínkek ...

Vyřeště optimalizační problém.

Protože hledáme afinní podprostor, je nutné nejprve body posunout tak, aby jejich těžiště $\bar{\mathbf{a}}$ bylo v bodě \mathbf{o} . Označme matici **A** jako:

$$\mathbf{A} = \left[\mathbf{a}_i - \bar{\mathbf{a}}, ..., \mathbf{a}_{25} - \bar{\mathbf{a}}\right]^T$$

kde $\bar{\mathbf{a}} = \frac{1}{25} \sum_{i=1}^{25} \mathbf{a}_i$.

Nyní řádkové vektory matice \mathbf{A} reprezentují body, které proložíme lineárním podprostorem dimenze 2. Víme, že ve spektrálním rozkladu matice $\mathbf{A}^T \mathbf{A} = \mathbf{V} \Lambda \mathbf{V}^T$ dva sloupce matice \mathbf{V} odpovídající největším vlastním číslům tvoří ortonormální bázi námi hledaného lineárního podprostoru (minimalizujícího součet čtverců vzdáleností k bodům \mathbf{a}_i).

Zároveň víme, že singulární čísla matice A (z SVD rozkladu) jsou druhé odmocniny vlastních čísel matice A^TA $(i \mathbf{A} \mathbf{A}^T)$ a singulární vektory SVD rozkladu matice \mathbf{A} odpovídají vlastním vektorům spektrálního rozkladu matice $\mathbf{A}^T \mathbf{A}$, protože

$$\mathbf{A}^T \mathbf{A} = \mathbf{V} \mathbf{S} \mathbf{U}^T \mathbf{U} \mathbf{S} \mathbf{V}^T = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T = \mathbf{V} \Lambda \mathbf{V}^T$$

Provedeme SVD rozklad matice A:

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

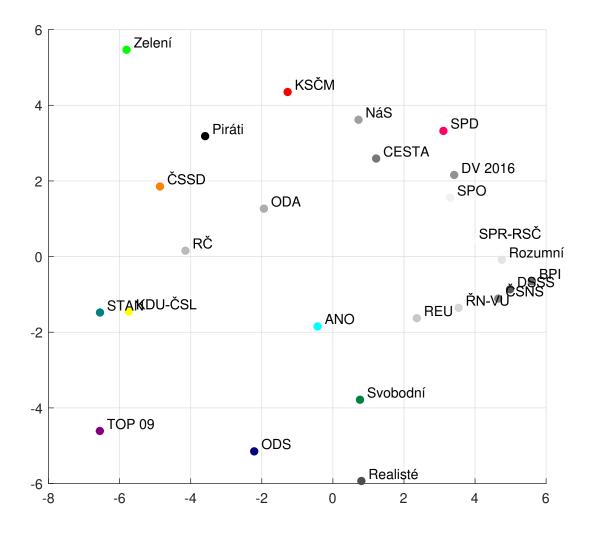
Druhá odmocnina je rostoucí funkce, proto vybíráme 2 největší singulární čísla matice A a jim příslušné vlastní vektory \mathbf{v}_1 , \mathbf{v}_2 z matice \mathbf{V} . Tyto vektory tvoří ortonormální bázi hledaného lineárního podprostoru.

Nyní body posunuté body transformujeme do nalezeného podprostoru. Protože je nalezená báze ortonormální, souřadnice v této bázi jsou koeficienty lineární kombinace těchto bázových vektorů rovné posunutým bodům vyjádřených v bázi kanonické.

$$\mathbf{B} = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix} \cdot \mathbf{A}^T$$

Sloupce matice ${\bf B}$ nyní obsahují souřadnice bodů ${\bf a}_i$ v nalezené bázi.

3. Najděte ortonormální bázi nalezeného afinního podprostoru span $\{a'_1,...,a'_{25}\}+a_0$ a jednotlivé politické strany zobrazte v souřadnicích této báze (jejíž vektory ztotožníte s osami 2D grafu). Zakreslené body obarvěte barvou strany.



pozn. opravujícího: Chybí zde hodnota kritéria.

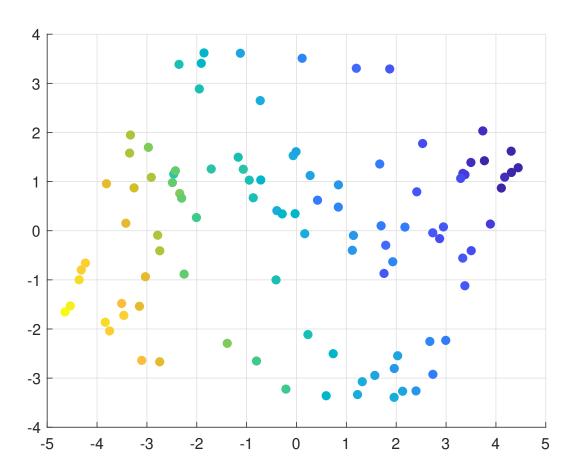
4. Interpretujte výsledek.

Jelikož nalezený afinní podprostor nejvěrohodněji prokládá odpovědi na otázky jednotlivých politických stran (tj. oba bázové vektory ukazují ve směrech největších diverzit odpovědí jednotlivých stran), z grafu můžeme vyčíst, jak jsou si politické strany navzájem názorově podobné - čím jsou si body reprezentující odpovědi daných stran blíže, tím podobnější názory tyto strany zastávají.

Úkol 2

Vizualizovat tato data je možné i z opačného pohledu. Jednotlivé otázky $\{\mathbf{b}_1,...,\mathbf{b}_{100}\}\in\mathbb{R}^{25}$ jsou body v prostoru politických stran. Otázky zobrazíme opět v prostoru dimenze 2.

1. Postupujte obdobně jako v minulém příkladu a zakreselete jednotlivé otázky v souřadnicích ortonormální báze dimenze 2 prostoru proložení ve smyslu nejmenších čtverců. Výsledné body obarvěte podle toho, zda většina politických stran odpovídá ano.



2. Interpretujte výsledek.

Tento graf narozdíl od předchozího zobrazuje informace o otázkách a jejich odpovědích politickými stranami (nikoli informace o samotných politických stranách, ty zde odpovídají jednotlivým dimenzím). Každá otázka má odstín podle četnosti odpovědi Ano. U modrých bodů (otázek) byla tato odpověd nejčetnější, u žlutých nejméně početná. Blízkost bodů v grafu obdobně jako minule reflektuje podobnost, tentokrát posloupnosti odpovědí na odpovídající otázku.

3. Je možné nalézt ortonormální bázi hledaného podprostoru a souřadnice promítnutých bodů v této bázi využitím výpočtu z předchozího úkolu? Pokud ano, vysvětlete.

Kdybychom hledali bázi lineárního podprostoru a nikoliv afinního (tj. těžiště $\bar{\mathbf{a}} = \bar{\mathbf{b}} = \mathbf{o}$), bylo by možné vybrat z matice \mathbf{U} (z SVD rozkladu matice \mathbf{A}) dva vektory \mathbf{u}_1 , \mathbf{u}_2 odpovídající největším singulárním číslům. Tyto vektory by tvořily ortonormální bázi hledaného podprostoru, protože

$$\mathbf{A}^T = \begin{bmatrix} \mathbf{a}_1, ..., \mathbf{a}_{25} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1, ..., \mathbf{b}_{100} \end{bmatrix}^T$$

$$\mathbf{A}^T = (\mathbf{U}\mathbf{S}\mathbf{V}^T)^T = \mathbf{V}\mathbf{S}\mathbf{U}^T$$

$$(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}\mathbf{S}\mathbf{U}^T = \mathbf{U}\mathbf{S}\mathbf{U}^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$$

Matice \mathbf{U} totiž obsahuje vlastní vektory spektrálního rozkladu $\mathbf{A}\mathbf{A}^T$.

Protože jsme ovšem nejprve u afinního prostoru museli provést posunutí prokládaných bodů odečtením těžiště, těžiště se liší pro body \mathbf{a}_i a \mathbf{b}_i (v matici se všemi daty počítáme průměr buď přes sloupce nebo přes řádky). Matice \mathbf{A} v těchto případech pak obsahuje odlišné hodnoty a další výpočty se pak liší.