Vlastné vektory a vlastné čísla v dátovej vede

Definícia vlastných vektorov a čísel

Nech A je matica typu $m \times m$. Nech $x \in R^m$ a $\lambda \in R$ spĺňajú

$$Ax = \lambda x.$$

Potom hovoríme, že x je vlastný vektor (*eigenvector*) matice A a λ je prislúchajúce vlastné číslo / vlastná hodnota (*eigenvalue*) matice A.

Základný pojem v matematike a fyzike, bohatá teória aj aplikácie (napr. geometrické transformácie, numerická matematika, diferenčné a diferenciálne rovnice, vo všeobecnejšej forme analýza vibrácií, kvantová mechanika)

Aplikácie vlastných vektorov a čísel v dátovej vede

- Numerické výpočty (často založené na príbuznom pojme „singulárnych hodnôt“ (*singular values*)),

- Markovské reťazce (napr. identifikácia „stacionárneho rozdelenia“ π a analýza rýchlosti konvergencie k π),

- Spektrálna analýza kovariančnej, korelačnej a informačnej matice v štatistike a v navrhovaní experimentov, ...

Pozitívne semidefinitné matice

Nech A je matica typu $m \times m$. Hovoríme, že A je pozitívne semidefinitná (*non-negative definite*), ak je symetrická a pre každé $x \in R^m$

$$x'Ax \geq 0.$$

Hovoríme, že A je pozitívne definitná (*positive definite*), ak je pozitívne semidefinitná a regulárna.

Každá kovariančná, korelačná a takzvaná informačná matica je pozitívne semidefinitná!

Špeciálne, pozitívne semidefinitná „výberová kovariančná matica“ (*sample covariance matrix*) pre realizáciu $x_1, \ldots, x_n \in R^m$ náhodného výberu je

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

kde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

je po (pozložkový) „výberový priemer“ (*sample mean*) vektorov dát $x_1, \ldots, x_n$.

Vlastné vektory a čísla pozitívne semidefinitnej matice

Dá sa ukázať, že ak je S pozítivne semidefinitná matica typu $m \times m$, tak existuje ortonormálny systém $u_1, \ldots, u_m$ vlastných vektorov matice S a

$$S = \sum_{j=1}^{m} \lambda_j u_j u_j',$$

kde $\lambda_1, \ldots, \lambda_m$ sú nezáporné vlastné čísla zodpovedajúce vlastným vektorom $u_1, \ldots, u_m$ (v uvedenom poradí). Iný zápis:

$$S = U \Lambda U',$$

kde $U = (u_1, \ldots, u_m)$ a $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. Rozklad matice S na takýto súčin niekedy nazývame spektrálny rozklad (*spectral decomposition*).

Ako súvisia vlastné čísla a vektory matice S s dátami?

Predpokladajme, že najmenšia rovina, v ktorej ležia všetky dátové body $x_1, \ldots, x_n \in R^m$, je $x_0 + L$, kde L je k-rozmerný lineárny vektorový priestor. Potom S má k nenulových vlastných čísel a vlastné vektory prislúchajúce týmto nenulovým vlastným číslam generujú L.

To znamená, že „v smeroch“ ktoré sú určené vlastnými vektormi, ktorým zodpovedajú nulové vlastné čísla, majú dáta „nulový rozptyl“.

Všeobecnejšie: Vlastné číslo λ prislúchajúce vlastnému vektoru u matice S určuje „rozptyl“ dát $x_1, \dots, x_n$ v smere vektora u.

Presnejšie: Ak u je vlastný vektor matice S, ktorému prislúcha vlastné číslo λ, tak „skóre“ $u'x_1, \dots, u'x_n$ majú výberový rozptyl $S_y^2 = \lambda$.* Skóre je možné chápať ako súradnice projekcie na priamku generovanú vektorom u.

Analýza/metóda hlavných komponentov (principal component analysis, PCA)

Porozprávame si o "empirickej" PCA, čiže takej, ktorá je založená priamo na dátach $x_1, \dots, x_n$ („teoretická“ PCA je založená na „populačnom rozdelení pravdepodobnosti“ náhodného vektora, ktorého sú $x_1, \dots, x_n$ realizácie).

1) Nájdeme ortonormálny systém vlastných vektorov $u_1, \dots, u_m$ matice S zoradený tak, aby pre zodpovedajúce vlastné čísla platilo

$$\lambda_1 \geq \cdots \geq \lambda_m.$$

2) Zvolíme vhodné k. Ak chceme pomocou PCA vizualizovať dáta, tak k je dimenzia vizualizačného zariadenia ($k = 2$, občas $k = 3$). Ak chceme pomocou PCA realizovať „redukciu dimenzie“, tak sa k volí tak, aby bol „dostatočne veľký“ „podiel vysvetleného rozptylu“

$$\alpha_k = \frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_k + \cdots + \lambda_n}$$

3) Centrované „skórové“ vektory (ktoré sú len k-rozmerné!)

$$y_1 = (u_1, \dots, u_k)'(x_1 - \bar{x}),$$

$$y_2 = (u_1, \dots, u_k)'(x_2 - \bar{x}),$$

$$\dots$$

$$y_n = (u_1, \dots, u_k)'(x_n - \bar{x})$$

použijeme na zobrazenie alebo na reprezentáciu dát $x_1, \dots, x_n$ v menšom (k-rozmernom) priestore.

Ukážme si príklad v R-ku.