RDD - 1
15 February 2021 14:44

samuel.rosa@fmph.uniba.sk

www.iam.fmph.uniba.sk/ospm/Rosa → všetko INFO

- odporúčam robiť si poznámky
- Hodnotenie: 70b - projekt
  : 30b - ústna skúška
  : detaily - stránka
  } raz, počas skúškového
- komunikácia je vítaná
- tykanie
- Literatúra: Izenman - Modern multivariate statistical techniques, 2008
  : Hastie et. al. - The elements of statistical learning, 2009

---

dáta:

	veľk	príjem	... ← premenné
osoba 1	29	1800	
osoba 2	62	913	
⋮	⋮	⋮	

↑ merania

všeobecne:

	1	2	...	k ← premenné
1		x_1^T		} x_i - malé
2		x_2^T		
⋮				
n		x_n^T		$= X$: $n \times k$

↑ merania

$x_i \in \mathbb{R}^k$, k môže byť veľmi veľké

- na zakreslenie
- zbytočné premenné v modeli (overfitting - pretrénovanie)
- príliš komplikované na pochopenie
- výpočtovo náročné

→ praktický / dátový pohľad
↳ $X_{n \times k}$, vektory $x_1, \dots, x_n$
→ teoretický / populačný
↳ x_i sú realizácie

náhodného vektora
$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ } X_j sú veľké

⇒ redukcia dimenzie dát

RDD

"dátový analytik" za 4 roky

"dátové predmety" v 4. roč. ⟨ RDD
⟨ Analýza veľkých dát ... (Harman)
↳ prebrané niektoré z metód z RDD
↳ odporúčam zapísať

RDD — a) kombinovanie starých premenných, aby sme vytvorili nové: PCA, faktorová analýza ... – extrakcia premenných
→ ①: lin. metódy, ②: nelineárne met. (feature extraction)
b) výber iba časti premenných – selekcia prem. (feature selection)
↳ Lasso, Hrebeňová regresia, ...
→ ③

– ale môže byť aj príliš veľké n – najmä numericky
→ konštrukcia podvzorky (subsampling) → ④ ak stihneme
: leveraging ...

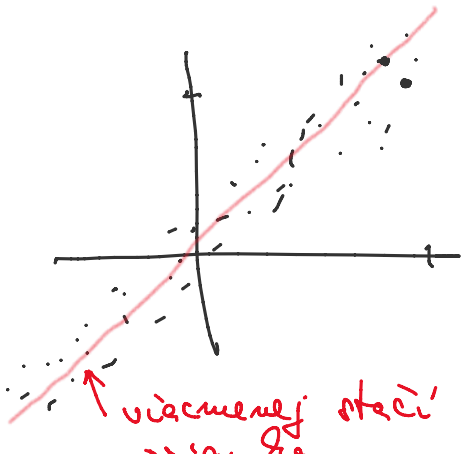
① Lineárne metódy extrakcie premenných [– lineárne v dátach]

Metóda hlavných komponentov (Principal component analysis – PCA)

$x_i \in \mathbb{R}^k$, chceme znížiť dimenziu
– ale nechceme stratiť priveľa informácie
– PCA: informáciu meria rozptýlenosťou

teoretický / populačný pohľad:

$X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}$ – náh. vektor



– kovariančná matica:

$$\Sigma = Cov(X) = E\left[(X - E(X))(X - E(X))^T\right]$$

– PCA pracuje so Σ → s jej spektrálnym rozkladom

$\underset{k \times k}{\Sigma}$ – r.s.d. $\Rightarrow$ $\underset{k \times k}{\Sigma} = \underset{k \times k}{U} \underset{k \times k}{\Lambda} \underset{k \times k}{U^T}$

$U^T U = I = U U^T$

U – ortogonálna, vl. vektory = stĺpce U

$p \times p$ $p \times p$ $p \times p$ $p \times p$

U - ortogonálna, vl. vektory = stĺpce U

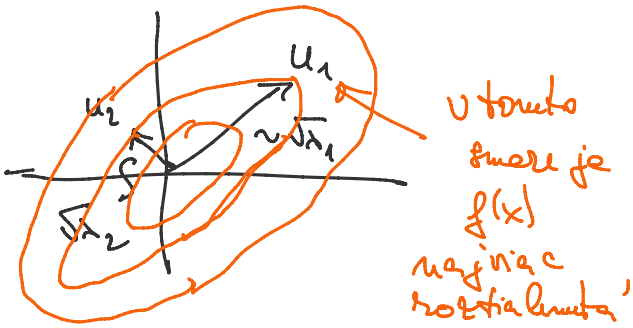
$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, λ_j - vl. hodnoty

$\Sigma u_j = \lambda_j u_j \quad \forall j = 1, \dots, p$ $\quad (Ax = \lambda x)$

Príklad: Ak $X \sim N_p(0, \Sigma)$, potom vrstevnice jeho hustoty sú elipsoidy so stredom v O_p a polosami v smeroch vl. vektorov matice Σ; dĺžky polosí $\sim \sqrt{\lambda_j}$ (↑ vl. hodnoty)

$$f(x) = \text{niečo} \cdot e^{-\frac{x^T \Sigma^{-1} x}{2}} = c$$

$$\Leftrightarrow x^T \Sigma^{-1} x = \text{konšt.}$$



Predp.: X: $E(X) = \mu$

$\mathrm{Cov}(X) = \Sigma = U \Lambda U^T$,

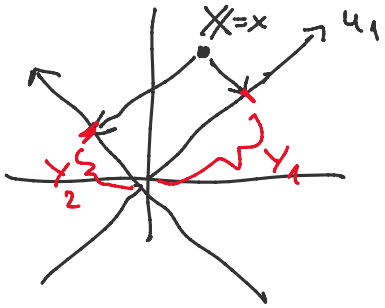
$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

- vezmeme súrad. systém daný vl. vektormi $u_1, \dots, u_p$ $\quad (U = (u_1, \dots, u_p))$
  - → to sú "kľúčové smery dát"
  - → aké sú súradnice X v súr. systéme danom U?

$$X = U \cdot Y \quad / \cdot U^T \quad \left(X = Y_1 \cdot u_1 + Y_2 \cdot u_2 \text{ pre } p = 2 \right)$$

$$\Rightarrow U^T X = Y$$

- toto je pre $\mu = 0$



pre všeob. μ: pracujeme s $X - \mu$:

$$X - \mu = U \cdot Y$$

$$\Rightarrow Y = U^T (X - \mu)$$: súradnice X vzhľadom na U

Y_j: dĺžka projekcie X na priamku danú u_j

Y_j: j-ty hlavný komponent

Definujme: $Y = U^T (X - \mu)$: vektor hlavných komponentov

$$E(Y) = U^T [\underbrace{E(X)}_{\mu} - \mu] = 0_p$$

$$\mathrm{Cov}(Y) = \mathrm{Cov}(U^T (X - \mu)) = U^T \mathrm{Cov}(X) U =$$

$$\boxed{\mathrm{Cov}(AX) = A \mathrm{Cov}(X) A^T}$$

μ

$$Cov(Y) = Cov(U^T(X-\mu)) = U^T Cov(X) U = U^T \Sigma U =$$

$\Sigma = U \Lambda U^T$

$$= \underbrace{U^T U}_{I} \Lambda \underbrace{U^T U}_{I} = \Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{bmatrix}_{k \times k}$$

$\Rightarrow$ Y_j sú nekorelované

$\Rightarrow D(Y_1) \geq D(Y_2) \geq \dots \geq D(Y_k)$

- $Cov(X, Y) = Cov(X, U^T(X-\mu)) =$

$$= Cov(X, U^T X) = Cov(X, X) \cdot U = U \Lambda U^T U = U \Lambda =$$

$\Sigma = U \Lambda U^T$

$$= \begin{bmatrix} | & | & & | \\ \lambda_1 u_1 & \lambda_2 u_2 & \dots & \lambda_k u_k \\ | & | & & | \end{bmatrix}$$

(i,j)-ty prvok U i-ty prvok vektora u_j

$$\Rightarrow cov(X_i, Y_j) = (\lambda_j u_j)_i = \lambda_j (U)_{ij} = \lambda_j (u_j)_i$$

$$\rho_{X_i Y_j} = \frac{\lambda_j (u_j)_i}{\sqrt{\Sigma_{ii} \cdot \lambda_j}} = \frac{(u_j)_i \sqrt{\lambda_j}}{\Sigma_{ii}}$$

$\rightarrow$ j-ty vl. vektor u_j určuje korelácie medzi Y_j a $X_1, \dots, X_k$.

$\uparrow$ j-ty hl. komp.

Veta: Prvý hl. komponent má najväčšiu disperziu spomedzi všetkých normovaných lin. kombinácií zložiek X. T. j.

$$D(Y_1) \geq D(l^T X) \quad \forall l \in \mathbb{R}^k : \|l\| = 1.$$

Dôkaz: $D(l^T X) = l^T \cdot Cov(X) \cdot l = l^T U \Lambda U^T l = v \Lambda v =$

$\Sigma = U \Lambda U^T$, $v^T = l^T U$, $v = U^T l$, $\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}$

$$= \lambda_1 v_1^2 + \dots + \lambda_k v_k^2 \leq \lambda_1 v_1^2 + \dots + \lambda_1 v_k^2 =$$

$\lambda_k \leq \lambda_1$

$$= \lambda_1 (v_1^2 + \dots + v_k^2) = \lambda_1 \|v\|^2 = \lambda_1 v^T v = \lambda_1 l^T U U^T l =$$

$$= \lambda_1 \underbrace{l^T l}_{1} = \lambda_1 = D(Y_1).$$

1, lebo $\|l\| = 1$ ▨

- druhý hl. komponent má najväčšiu disperziu spomedzi všetkých norm. lin. komb. na ortogonálnom komplemente u_1:

$D(Y_2) \geq D(l^T X) \quad \forall l \in \mathbb{R}^k : \|l\| = 1, \; l \perp u_1$ OPRAVA: $\leftarrow$ ($Y_1 = u_1^T (X - \mu)$.

$D(Y_2) \geq D(a^T X) \quad \forall a \in \mathbb{R}^p : \|a\| = 1,\ a \perp \cancel{X_1}\ u_1$

OPRAVA: ↙ ↘

$(Y_1 = u_1^T (X - \mu)$, čiže Y_1 získame pomocou $a = u_1$ ⇒ pre ďalšie Y_j musí byť $a \perp u_1$)

atď, pre Y_3: $\quad -\!\!\!\parallel\!\!\!- \quad a \perp \cancel{X_1}\ u_1,\ a \perp \cancel{Y_2}\ u_2$