projekt - zadanie na stránke — 2 metódy aplikované na dátach

- PDF text + prezentácia na skúške
- niečo zaujímavého sa dozvedieť

• na stránku dnes dám formulár - spätná väzba → napíšte, ak chcete niečo zmeniť

Minule: Proj. sledovanie

$q=1$: $Y = \ell^T \mathbb{X}$ — najzaujímavejší smer

$\varphi(Y) \to \max\limits_{\ell \in \mathbb{R}^q}$

prakticky : $\underset{n\times 1}{Y} = \underset{n\times q}{X}\,\underset{q\times 1}{\ell}$

$\varphi \to$ čo "najnenormálnejšie" dáta

φ - 1.) $\varphi(Y) = D(Y) \Rightarrow$ PCA

2.) $\varphi(Y) = |\mathcal{H}_4(Y)|$

$$\varphi(Y) = \frac{\mathcal{H}_3(Y)^2}{12} + \frac{\mathcal{H}_4(Y)^2}{48}$$

max φ : nekonvexná optimalizácia

praktické použitie :

- každé lok. max. - potenciálne zaujímavá proj.
- spustíme veľakrát, zapamätáme si všetky výsledky
- vyhľadáme výrazne rôzne projekcie < výrazné poklesy $\varphi(Y)$ / uhly medzi smermi ℓ

Dnes:

↳ napr. ak proj$_1$ a proj$_{10}$ sú výrazne iné, môžeme ich použiť na 2D reprezentáciu

Pr.: olive - zloženia olivových olejov

Ďalšie φ :

3.) Vzdialenosť od hustoty $N(\,)$:

$$\varphi(Y) = \int_{\mathbb{R}} \big(\underset{\uparrow\ \text{hustota } N(0,1)}{f_N(y)} - \underset{\uparrow\ \text{hustota } Y}{f_Y(y)}\big)^2 \underset{\uparrow\ \text{váhy}}{w(y)}\,dy \to \max$$

• $w(y) \equiv 1 \Rightarrow \varphi(Y) \doteq \Big[E(f_N(Y)) - \underbrace{E(f_N(Z))}_{=\frac{1}{2\sqrt{\pi}}}\Big]^2$, kde $Z \sim N(0,1)$

↑ rovnaké váhy na všetko

• $w(y) = \dfrac{1}{f_N(y)}$: čím ďalej od stredu, tým väčšia váha

$$\Rightarrow \varphi(Y) \doteq \frac{(\mathcal{H}_3(Y))^2}{12} + \frac{(\mathcal{H}_4(Y))^2}{48} \to 2.)$$

- $w(y) = f_N(y)$: najviac váhy do stredu

$\Rightarrow \varphi(Y) \doteq \left[E(f_N(Y)) - E(f_N(Z)) \right]^2$: to isté ako pri $w = 1$

prakticky: $\hat{\varphi}(y)$ $\longrightarrow$ cez aproximácie a potom: $\hat{\varkappa}_j$, $\hat{E}$

$\searrow$ aproximácia pomocou polynómov

4.) Shannonova entropia (SHANNON)

↳ miera neusporiadanosti

↳ $I(Y) = -\int_{\mathbb{R}} f_Y(y) \ln f_Y(y)\, dy$

- pre náh. veličiny s rovnakou varianciou: $I(Y)$ je maximálna pre $N()$

my chceme čo najmenej normálne: $\varphi(Y) = 0$ pre $N()$

$\varphi(Y) \geq 0$ pre ~~NETS~~

$\Rightarrow \varphi(Y) = \underbrace{I(Z) - I(Y)}_{\to \max}$, kde $Z \sim N(0,1)$

$I(Z) \doteq \left\langle \begin{array}{l} + \frac{1}{2}(1 + \ln(2\pi)) \doteq 1{,}419 \\ \cancel{-} \end{array} \right.$

↑ overím : overené

v praxi: $\tilde{\varphi}(Y) = -I(Y) \to \max$

odhad: $\rightarrow \hat{\varphi}(y) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_Y(y_i) \cdot \ln \hat{f}_Y(y_i)$

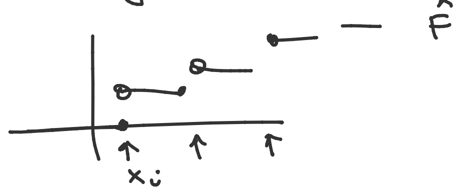
↳ $\hat{f}$: jadrový odhad hustoty

$\searrow$ aproximácia pomocou momentov: $\varphi(Y) \doteq \frac{(\varkappa_3(Y))^2}{12} + \frac{(\varkappa_4(Y))^2}{48}$

$\rightarrow \hat{\varkappa}_j$

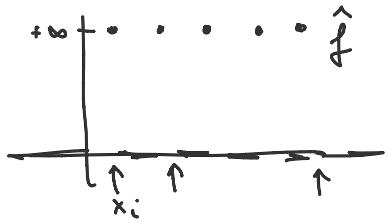
Odbočka: jadrový odhad hustoty (kernel density estimation)

dáta $x_1, \dots, x_n$ $\longrightarrow$ $\hat{F}(x)$: schodikovitá f.
zo spojitého rozd.



- odhad rozdelenia – histogram

hustota? – $\hat{F}$ zodpovedá hustote:

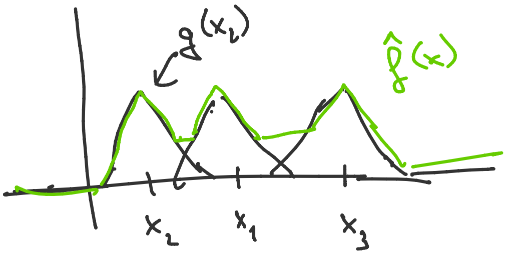


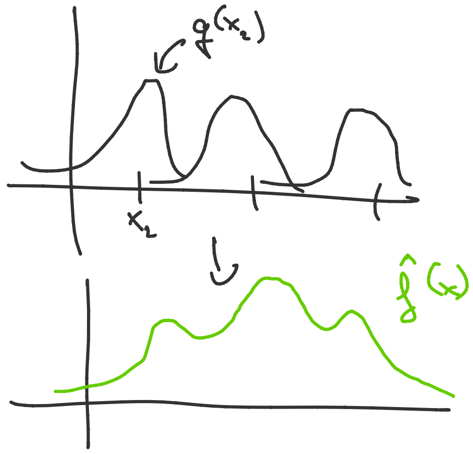
$\hat{f}(x) = \frac{1}{n} \sum \delta_{x_i}(x)$ $\rightarrow$ ZLÉ

$\delta_t(x) = \begin{cases} +\infty, & x = t \\ 0 & \text{inak} \end{cases}$

jadrové odhady: v bodoch x_i nahradíme $\delta_.$ kopčekmi so stredmi v x.

jadrové odhady: v bodoch x_i nahradíme δ_{x_i} kopčekmi so stredmi v x_i





$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{g_{x_i}(x)}_{\text{jadro}}$$

bandwidth h : naškálovanie

: pre $h \to 0$: veľmi úzke a vysoké kopčeky

formálne : $$g_{x_i}(x) = \frac{1}{h}\, \underbrace{\text{hustota}}_{\text{jadro}}\left(\frac{x - x_i}{h}\right)$$

Viacrozmer : $q > 1$

a) • najprv 1-D projekcia pomocou h_1^*

- potom transformovať dáta do $\perp h_1^*$ (projektovať na $\mathcal{C}(h_1^*)^\perp$)
- zase 1-D projekcia transformovaných dát

b) $q = 1$: $\varphi(h^T \mathbb{X})$, $h \in \mathbb{R}^k$, $\|h\| = 1$

$q > 1$: $\varphi(\underset{q\times k}{H^T}\underset{k\times 1}{\mathbb{X}})$, $h_i \in \mathbb{R}^k$, $\|h_i\| = 1$

$H_{k\times q} = (h_1 \dots h_q)$ $\quad h_i \perp h_j \ \forall i \neq j$ $\Big\}$ $H^T H = I_q$

$$\max_{\substack{H \in \mathbb{R}^{k\times q} \\ H^T H = I_q}} \varphi(H^T \mathbb{X}) \quad : \text{teoreticky}$$

výberovo: $y_i = H^T x_i$

$$\underbrace{\begin{pmatrix} | & & | \\ y_1 & \cdots & y_n \\ | & & | \end{pmatrix}}_{Y^T} = H^T \underbrace{\begin{pmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{pmatrix}}_{X^T} \quad : \quad \underset{n\times q}{Y} = \underset{n\times k}{X}\ \underset{k\times q}{H}$$

→ potrebujeme verziu φ pre viacrozmerné Y

– nie vždy ex., nie vždy sa jednoducho rozšíri

– napr. entropia : $\varphi(Y) = \iint f(y_1, y_2) \ln f(y_1, y_2)\, dy_1 dy_2$

- napr. <u>entropia</u> : $\varphi(Y) = \iint_{\mathbb{R}^2} f_Y(y_1, y_2) \ln f_Y(y_1, y_2)\, dy_1 dy_2$

  $(q=2)$

  $\rightarrow \hat{\varphi}(Y) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(\underbrace{y_i}_{\in \mathbb{R}^2}) \ln \hat{f}(y_i)$ — výpočtovo náročné

- <u>momenty</u> : $\exists$ 2-rozmerné verzie

<u>Ďalšie viacrozmerné indexy</u> $(q > 1)$

<u>1. central mass index</u> : $\hat{\varphi}(Y) = \dfrac{\frac{1}{n} \sum_{i=1}^{n} e^{-\frac{1}{2} y_i^T y_i} - e^{-\frac{q}{2}}}{1 - e^{-q/2}}$ — vyhľadáva projekcie s veľa dátami v strede

<u>2. holes index</u> : $\hat{\varphi}(Y) = \dfrac{1 - \frac{1}{n} \sum_{i=1}^{n} e^{-\frac{1}{2} y_i^T y_i}}{1 - e^{-q/2}}$ — vyhľ. proj., ktoré majú v strede dieru

$\rightarrow$ znova vyšli z hľadania nenormality