

ju?

Minule: Princípy štatistiky

Bootstrap → "simulačné" výpočty generovaním z dát

- funguje aj všeobecne (iné dáta, iné modely)
 - niekedy treba modifikovať
 - nie vždy rozhodujú

Typy dát:

- kvantitatívne (výška, vek) - klasika
- kvalitatívne (kvalitatívne) - farba (biela, červená, ...)
 - majina (SK, CZ, ...)

...

↳ x-ka: typický pre kódovanie: one-hot-encoding, dummy premenne, ...

↳ y-ka: iné metódy: klasifikácia

(logistická reg., rozhodovacie stromy, ...)

- ordinálne: usporiadané kategórie - odpovede v dotazníku (maľo, stredne, veľmi, veľmi)
- zakategorizovanie spojitéj:

vek $\begin{cases} \in [0, 18] \rightarrow 1 \\ \in (18, 25] \rightarrow 2 \\ \in (25, 35] \rightarrow 3 \\ \vdots \end{cases}$

⇒ typický problém: (i) trváme sa, že je kvantitatívna (ii) pracujeme ako s kvalitatívnou

napr. na lin. reg.:

$$(i) \quad y_i = \beta_0 + \beta_1 (\text{veková kategória})_i + \varepsilon_i$$

1 2 3 4 ...

vek < 18 : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$
 vek ∈ (18, 25] : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 2$
 ... : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot 3$

} posun o 1 kategória
 ⇒ vždy rovnaký posun / pokles

$$\begin{aligned}
 \text{vek} \in (18, 25] : \hat{y}_i &= \beta_0 + \beta_1 \cdot 2 \\
 (25, 35] : \hat{y}_i &= \beta_0 + \beta_1 \cdot 3 \\
 &\vdots
 \end{aligned}
 \quad \Rightarrow \text{vždy rovnaký nárast / pokles}$$

: silné obmedzenie
(iba 1 parameter pre vek)

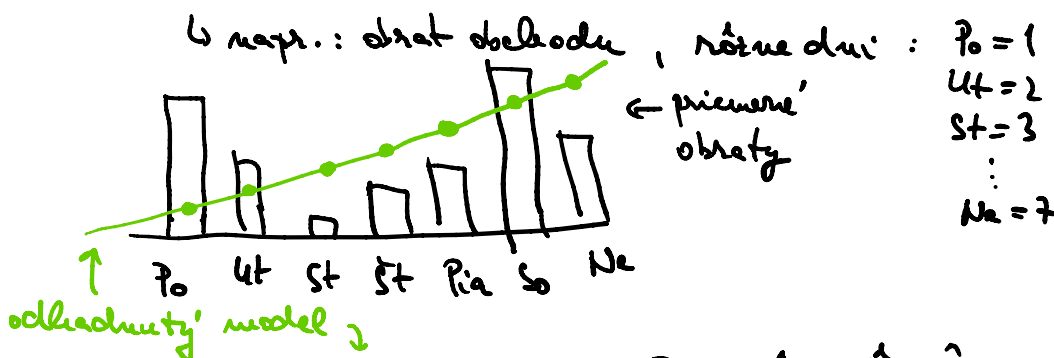
$$\begin{aligned}
 \text{(ii): } \text{kat}1_i &= \begin{cases} 1 \\ 0 \end{cases}, \quad i - \text{ty členok : vek} < 18 \\
 \text{kat}2_i &= \begin{cases} 1 \\ 0 \end{cases} \quad : \text{vek} \in (18, 25] \\
 &\vdots
 \end{aligned}$$

$$\Rightarrow y_i = \beta_0 + \beta_1 \text{kat}1_i + \beta_2 \text{kat}2_i + \beta_3 \text{kat}3_i + \dots + \varepsilon_i$$

$$\begin{aligned}
 \text{pre vek} < 18 : \hat{y}_i &= \hat{\beta}_0 \\
 \in (18, 25] : \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \\
 (25, 35] : \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_2 \\
 &\vdots
 \end{aligned}
 \quad \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \begin{array}{l} \text{každá veková kategória} \\ \text{má iný odhadovaný } y \\ \\ \text{: veľkosť vôle modelu} \\ \text{(1 parameter pre každú} \\ \text{kategóriu okrem jednotky)} \end{array}$$

všobecné: kategórie x_j v modeli $\Rightarrow$ veľa parametrov, zložitejší model
: kvantitatívne $x_j \Rightarrow$ menej parametrov (typicky 1), jednoduchší model

Pozn: kategórie sú niekedy zapísané ako čísla



$$y_i = \beta_0 + \beta_1 \text{den}_i$$

$\begin{matrix} 1 & 2 & 3 & \dots \end{matrix}$

$$\begin{aligned}
 \text{Po : } \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \\
 \text{Ut : } \hat{y}_i &= \hat{\beta}_0 + 2\hat{\beta}_1 \\
 &\vdots
 \end{aligned}
 \quad \left. \begin{array}{l} \\ \\ \end{array} \right\} \begin{array}{l} \text{lineárny} \\ \text{nárast / pokles} \end{array}$$

- typicky s kateg. prem. chceme robiť ako ozaj s kategóriou
 $\Rightarrow$ prehodovať

$$\Rightarrow y_i = \beta_0 + \beta_1 \text{ut}_i + \beta_2 \text{st}_i + \dots$$

$$\Rightarrow y_i = \beta_0 + \beta_1 \underset{\substack{\uparrow \\ 0}}{u_{t,i}} + \beta_2 \underset{\substack{\uparrow \\ 0}}{st_{t,i}} \dots$$

- nielen v lin. reg.

Klasifikácia

- predikcia správneho označenia (class, label) dátového bodu
 $\hookrightarrow$ z diskrétnych tried

Strojové učenie - INF, Analýza zhlukov a klas. dát - PHD

Binárna klasif. : 2 triedy : $\text{HuĽ}/\text{Ľena}$
 : $\text{banĽerot}/\text{nebanĽerot}$

Dáta:

$$X = \begin{bmatrix} \text{---} x_1^T \text{---} \\ \vdots \\ \text{---} x_n^T \text{---} \end{bmatrix}_{n \times p}$$

dátove body

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1}$$

: cieľová prem. (target), značka (label)...

$y_i \in \{-1, +1\}$, alebo $\{0, 1\}$ apod.

Klasifikač. reg. : $y_i \in \{0, 1\}$, $\hat{y}_i \in \mathbb{R}$

$\rightarrow$ dalo by sa : $x_i^T \hat{\beta} > 0.5 \Rightarrow \hat{y}_i = 1$

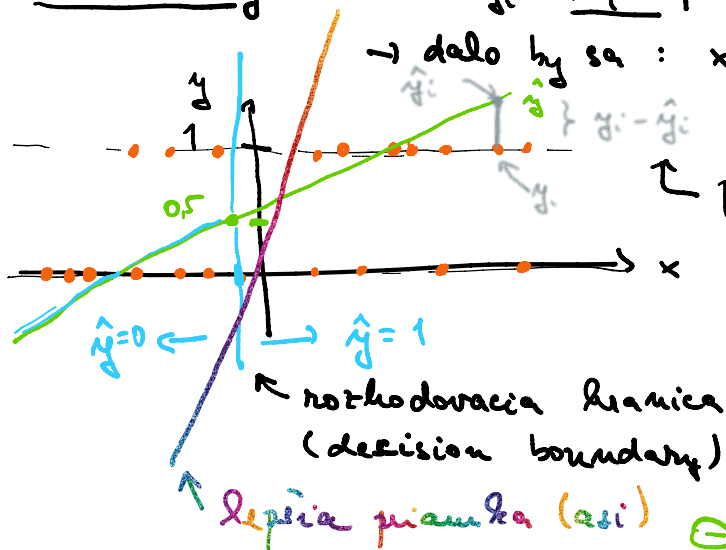
$< 0.5 \Rightarrow \hat{y}_i = 0$

$\hookrightarrow$ penalizujeme za to, že je príliš jasne 1 (lebo $y_i = 1$, $\hat{y}_i = 1.8$) : penalizujeme aj dobre rozhodnutia

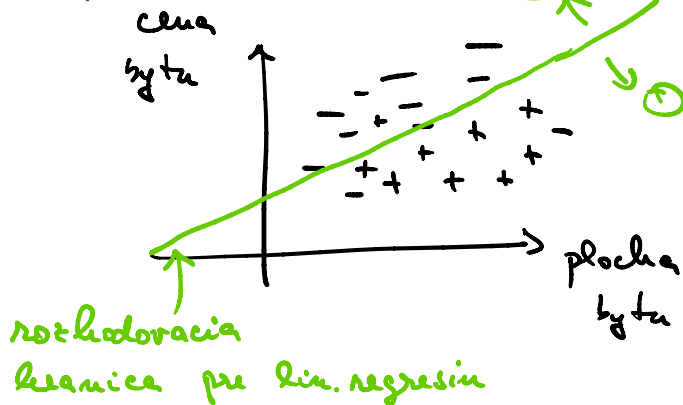
$\hat{y} = 0$ $\hat{y} = 1$

NIE IDEÁLNA METÓDA

+ : zaujímavá ponuka
 - : učite nezáujem



dve x :



k-Nearest Neighbors (k-NN) : k najbližších susedov

$k=1$: podľa najbližšieho suseda

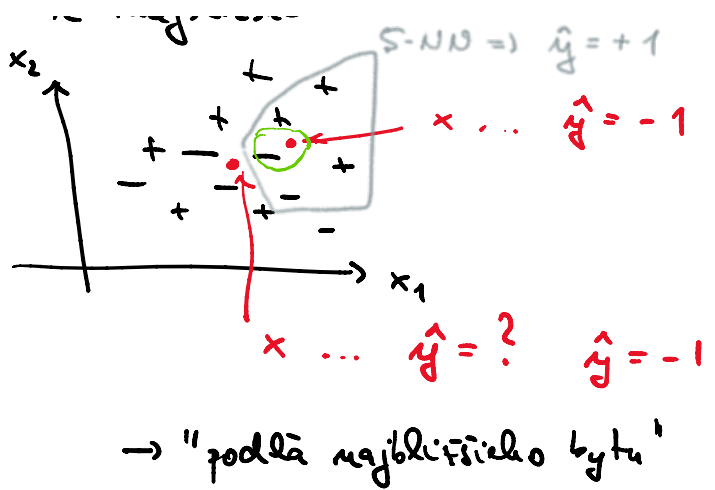
$x_2 \uparrow$

$5\text{-NN} \Rightarrow \hat{y} = +1$

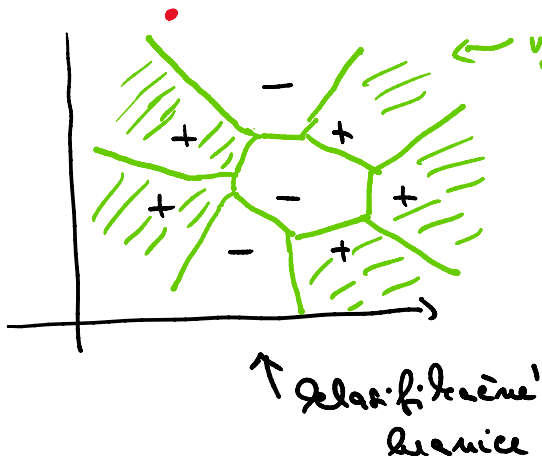
k - nearest neighbors

k=1: podľa najbližšieho suseda

- máme nové $x \in \mathbb{R}^p$
- nájdeme najbližšieho data
... $i^* = \arg \min_{i=1, \dots, n} \|x - x_i\|$
- zober jeho značku
... $\hat{y} = y_{i^*}$



všetchno princíp metód: Myšlienka $\rightarrow$ Matematika $\rightarrow$ Realizácia



← vyšetrovanie: predikujeme ⊕

} možný najbližší bodov:
Voronoi diagram

- najjednoduchší postup
- funguje niekedy prekvapivo dobre
- komplikovaná decision boundary $\Rightarrow$ nízke pretečovanie
- model s "veľkou voľnosťou"

$\Rightarrow$ stabilita: k-NN: k najbližších susedov

$\hookrightarrow \hat{y}$ = značka, ktorej je medzi "k" susedmi najviac
(hlasovanie)

- k: hyperparameter $\rightarrow$ voľba (cross-validation)
- outlieri: ťažšie ich dobre klasifikovať
: vo vstupnom datasete výrazne ovplyvňujú oblasť oblasť
- časová náročnosť: pre 1 predikciu treba potenciálne všetky vzdialenosti: $O(n^2)$ operácií
 $\rightarrow \exists$ efektívnejšie spôsoby
- algoritmy pre Vor. diagramy (iba o veľkosti "n")

→ $\exists$ efektywniejsze sposoby

- algorytmy na Vor. diagramy (iba przy małej "p")

- spec. algorytmy - rozkładanie przestrzeni na nury

- spec. strony

...