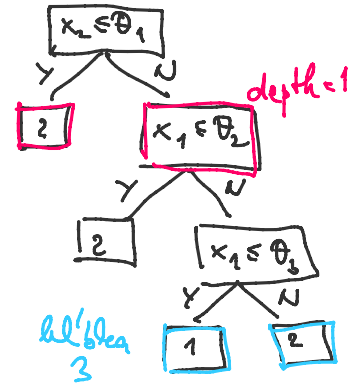
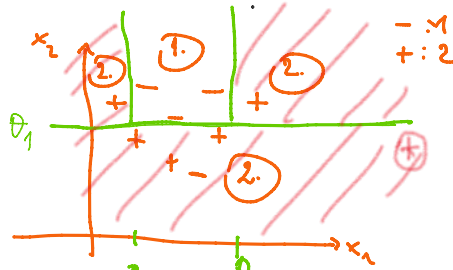


18.11., 14:10, B - pís.

Miniele: Rozhodovacie stromy



- hl'ba (uzlu) (depth): # rozhodovacích uzlov pred nim

Ako volit podmienky do uzlov

- z dát: trenovacia sada

ciele: čo najmenšia dĺžka

• nie príliš zložitý strom

↳ nie priveľa uzlov

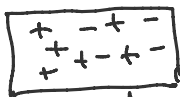
↳ aby sa neopreteroval

} bias-variance tradeoff

postup: teoria model

: v koncovom uzle: predpoveď najpočetnejšie triedy

: v uzle



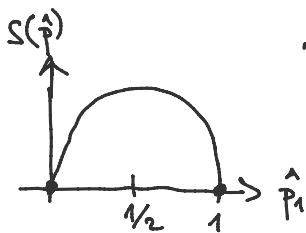
$$\begin{aligned} &\rightarrow \begin{matrix} -: 4 \\ +: 6 \end{matrix} \Rightarrow \begin{aligned} \hat{p}_1 &= \frac{4}{10} \\ \hat{p}_2 &= \frac{6}{10} \end{aligned} \end{aligned}$$

ideál:



- vyberáme premennú a podmienku, kt. "najlepšie rozdelí" → najviac vyčistí
pre $C=2$: ($x_5 \leq 7$)

čistota uzlu: miera chaosu: Shannonova entropia $= -(\hat{p}_1 \ln \hat{p}_1 + \hat{p}_2 \ln \hat{p}_2) = S(\hat{p})$



$$\hat{p}_i = \frac{\# \text{ dát so značkou } i}{\# \text{ dát celkovo}} \leftarrow \text{v danom uzle}$$

(podobne pre $C > 2$)

... $S(\hat{p})$ je maximálna, keď $\frac{1}{2}$ je - a $\frac{1}{2}$ je +

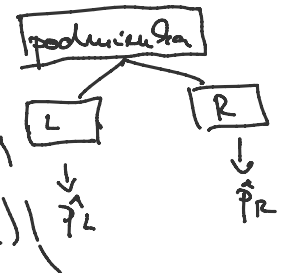
↳ "úplný chaos"

↳ "najmenej čistý uzol"

⇒ premenná a hranica, aby "nové" $S(\hat{p}) \rightarrow \min$

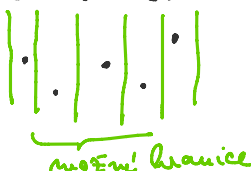
$$\Rightarrow \text{skóre} = S(\hat{p}_L) \cdot \frac{|L|}{|L|+|R|} + S(\hat{p}_R) \cdot \frac{|R|}{|L|+|R|}$$

: vážený
priemer
(váhy podľa
dát v L, R)



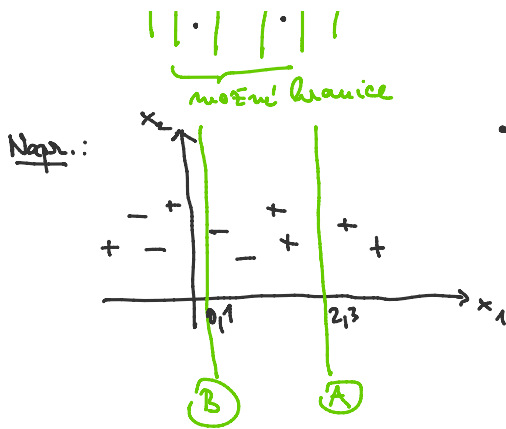
• premenná: konečný #

• hranice: x - keď diskretizujeme (resp. stačí $n+1$ hraníc v každej súradnici)



Konečne veľa
podmienok

podmienky



• vyberáme z (A) a (B)

najprv: $- : 4 \Rightarrow \hat{p}_1 = \frac{4}{10} = 0,4$
 $+ : 6 \Rightarrow \hat{p}_2 = \frac{6}{10}$

$\Rightarrow S = -(0,4 \ln(0,4) + 0,6 \ln(0,6)) = 0,67$

(A) $L: - : 4 \Rightarrow \hat{p}_1 = \hat{p}_2 = \frac{1}{2}$
 $+ : 4$

$\Rightarrow S(\hat{p}_L) = -\left(\frac{1}{2} \ln \frac{1}{2} + \frac{1}{2} \ln \frac{1}{2}\right) = 0,69$

$R: - : 0 \Rightarrow \hat{p}_1 = 0$
 $+ : 2 \Rightarrow \hat{p}_2 = 1 \Rightarrow S(\hat{p}_R) = 0$

$\Rightarrow \text{skóre} = \frac{8}{10} \cdot 0,69 + \frac{2}{10} \cdot 0 = 0,55$

(B) ... skóre = 0,65 $\Rightarrow$ zvolili by sme (A)

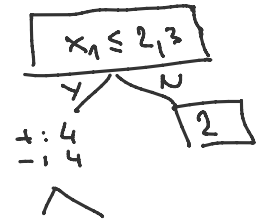
- namiesto entropie napr. Gini index $\rightarrow$ information gain
- podmienka zastavenia: napr. ak podľa entropie je už malý
 (v príklade $0,67 - 0,55 = 0,12$)

↳ alternatíva: ak by sme všetky uzly nie sú plne čisté (pure)
 ↳ iba 1 trieda

- príliš veľký strom

$\Rightarrow$ spätné škrcovanie (pruning)

- ďalšie sa aj na regresné problémy: regresné stromy
 (regression trees)



výhody: - interpretovateľné, ľahko vysvetliteľné

nevýhody: - nerobustné: malá zmena dát môže spôsobiť veľkú zmenu stromu
 - typicky slabšie predikcie (môžu overfitať)

Sady stromov

- kolektív (ensembles) stromov

1.) Bagging: veľa stromov

: $\{x_i\} \in \mathbb{R}^p$ priradiť taký label, na ktorom sa hodne najviac stromov

$\rightarrow$ ako ich získame? Bootstrapovať train set:

Train: $(x_1, y_1), \dots, (x_n, y_n)$

náhodne s návratom vyberiem n -krát (x_i, y_i)

$\rightarrow (x_1^*, y_1^*), \dots, (x_n^*, y_n^*) \rightarrow$ nafitujeme strom 1

$\rightarrow$ zopakujeme $\rightarrow$ strom 2

→ opakuje sa → strom 1

⋮

→ strom B (napr. $B = 1000$)

• rozl. stromy: problém s veľkou variáciou

• priemer dát: znižuje variáciu ($D(x_1) = \dots = D(x_n) = \sigma^2 \Rightarrow D(\bar{X}) = \frac{\sigma^2}{n}$)
metóda

⇒ Bagging (klasovanie & priemer): znižuje var ⇒ znižuje overfitting

2.) Náhodné lesy: bagging - stromy sú (silne) kontrované

- ak náhodná prem. je kľúčová, skoro všetky je na zao.

výbere ⇒ vyberajú podobne

⇒ nemusia dôjsť k takému zníženiu var

⇒ lesy: ako Bagging, ale v každom uzle pri výbere splitu zvážime iba náhodnú podmnožinu prem. ⇒ stromy sa budú viac líšiť

3.) Boosting: postupná tvorba stromov
: "postupné učenie":

1.) začneme so slabým stromom (limitovaná max. hĺbka)

2.) ďalší strom - na tých istých dátach

- znova slabý

- posilníme náhodnú podm. 1. strom chybne klasifikoval → Adaboost

↳ alternatíva: fituje reziduál predch. stromu → XG Boost, gradient boosting

→ opakuje sa 2 krát

- výsledok: klasovanie

- môže overfitovať, obzvlášť ak v Train sú niektoré dáta zle označené

Výhody ensembles: lepšie predikujú, menej overfitujú

Nevýhody: strátame interpretovateľnosť

ďalšie met. klasifikácie: Lineárna diskriminačná analýza, Support vector machines, ...

Strojové učenie - ML

AI & DL - ML