Strojové učenie a neurónové siete - **zadanie 3** - súborové učenie

29. október 2020
I-SUNS cv. 6

Zadanie 3

- Zadanie a bodovanie je k dispozícii na dokumentovom serveri v AIS
- Máte sa naučiť:
  1. Trénovať súborové učenie
  2. Zopakovať si klasifikáciu a regresiu
- Dataset:
  - dostupný v AIS aj s popismi stĺpcov
  - testovacia množina je oddelená, aby sa dali systémy navzájom porovnávať

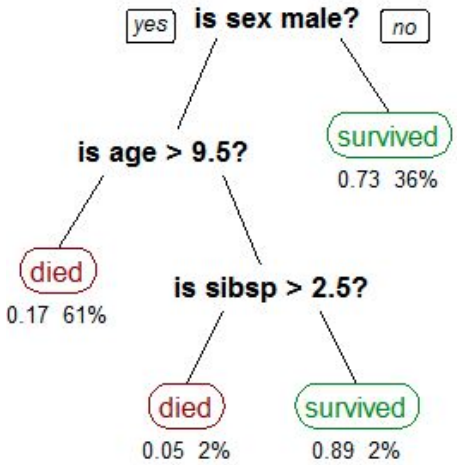
Obsah tejto prezentácie

- Rozhodovacie stromy
- Súborové učenie

- Ukážka - sklearn, plotly, pandas
- O týždeň: Bayesov naivný klasifikátor, RBF neurónové siete, zopakovanie regresie a vyhodnocovania

Rozhodovacie stromy

- Učenie s učiteľom
- Regresia aj klasifikácia
- Pekné:
  - Potrebuje málo predspracovania vstupov (nemusíme ani normalizovať)
  - Ľahko interpretovateľný - vyplývajú z neho if-else pravidlá
  - Zvláda multi-output problémy
- Menej pekné:
  - Ťažšie generalizuje (nevidené dáta; pretrénovanie)
  - Nestabilný (citlivé na vstupné dáta)
  - Ťažko sa hľadá optimum (heuristické algoritmy)

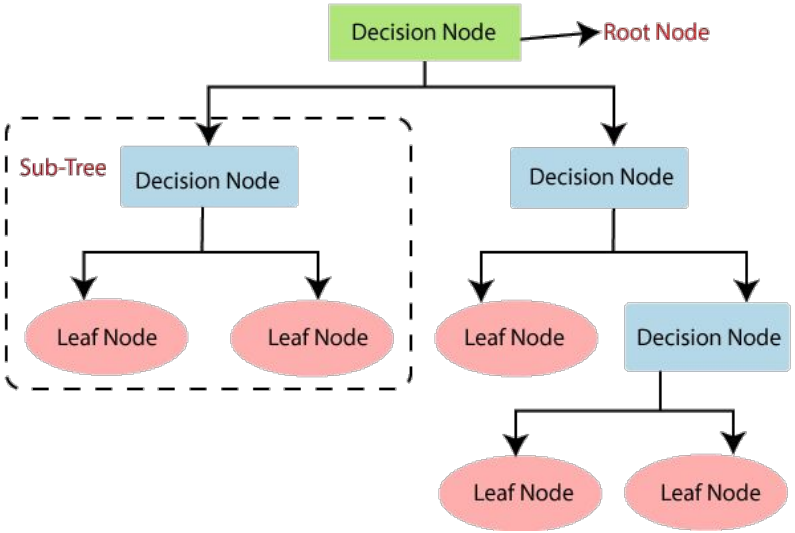


Zdroje obrázkov:
https://en.wikipedia.org/wiki/Decision_tree_learning

Rozhodovacie stromy - súčasti

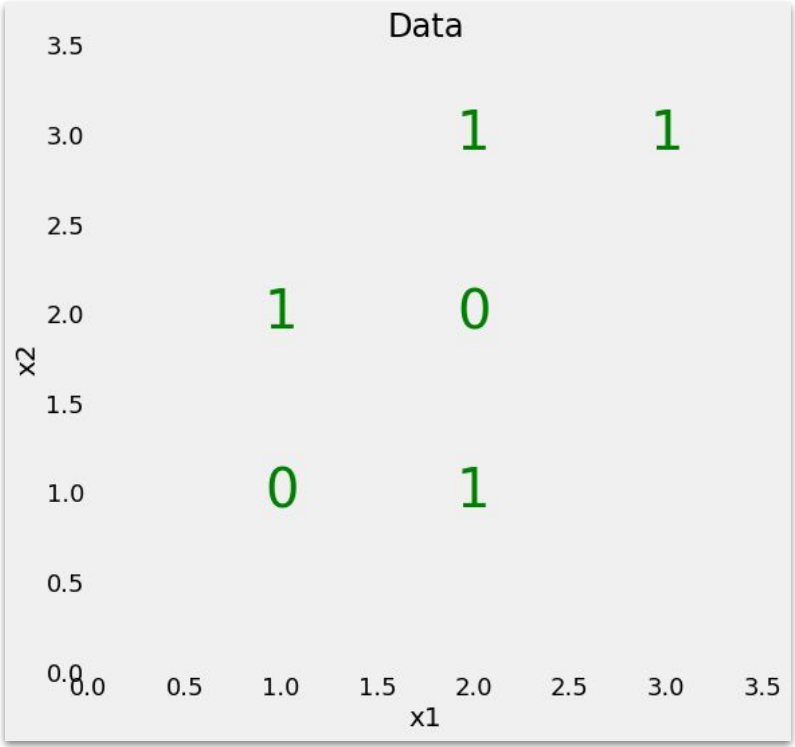
- Koreň (*root*) - obvykle sa kreslí navrchu
- Vnútorný uzol (*internal node*) - podmienka
- List (*terminal node, leaf, edge*) - rozhodnutie (výstup)

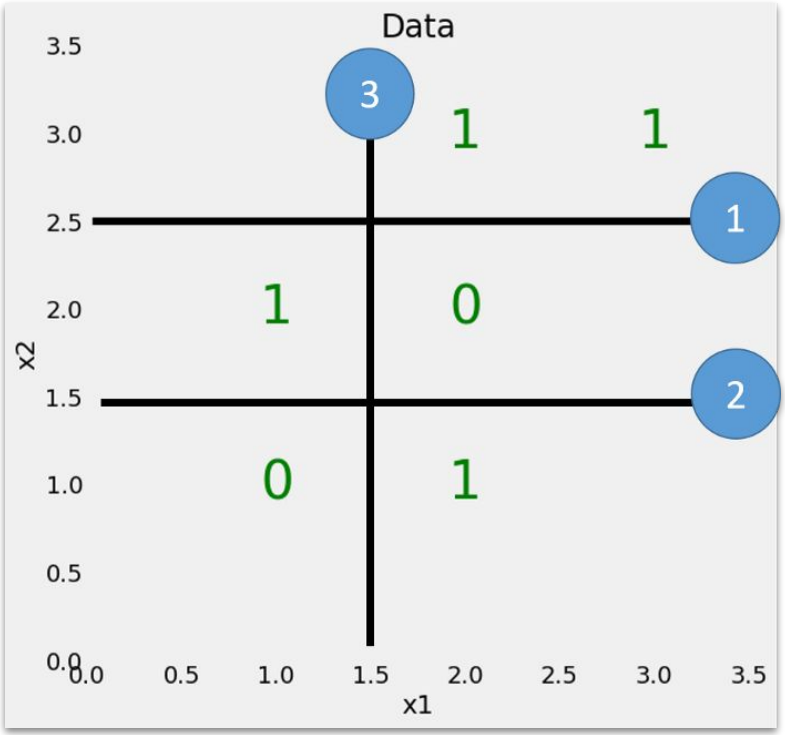
- Čím viac vstupných príznakov, tým väčší strom
- Vytváranie stromu - *induction* a *pruning*
- Algoritmy na vytváranie stromu:
  - CART
  - ID3, C4.5



Zdroje obrázkov:
https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

Rozhodovacie stromy

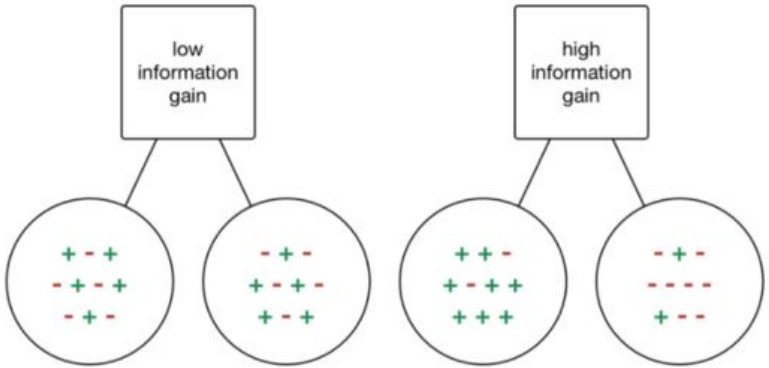




Rozhodovacie stromy - *induction*

1. Nájdi najlepší príznak a hranicu, na ktorých rozdeliť dáta
2. Rozdeľ podľa podmienky trénovacie dáta
3. Rekurzívne späť na krok 1., kým sa nesplní zastavovacia podmienka (presnosť, počet vzoriek v listoch, počet uzlov)

- Ako nájsť najlepší príznak na rozdelenie dát?



Zdroje obrázkov:
https://towardsdatascience.com/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956

Rozhodovacie stromy - *induction*

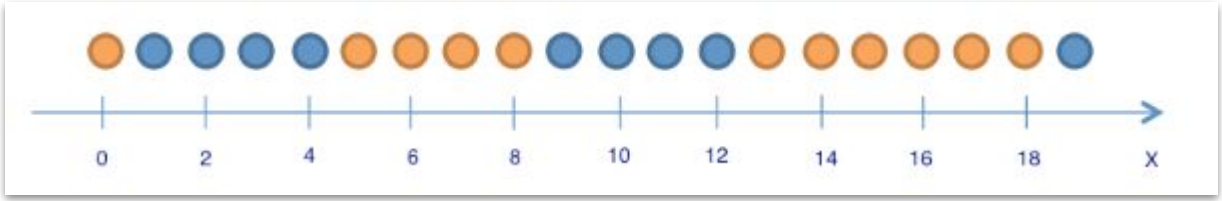
1. Vybranie hranice:
   - Diskrétne hodnoty - skúšam každú
   - Spojité hodnoty - snažím sa prehľadať priestor
2. Evaluácia metriky:
   - Klasifikácia:
     - Gini (CART)

$$gini = 1 - \sum_{i=0}^{c}(p_i)^2$$

     - Entrópia (ID3)

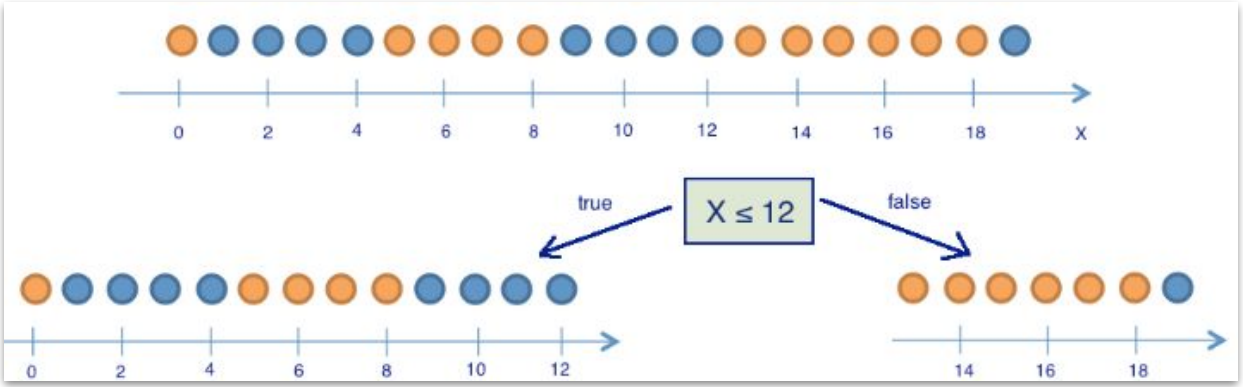
$$entropy = \sum_{i=1}^{C} -p_i * log_2(p_i)$$

   - Regresia:
     - MSE, reziduály



$$entropy = \sum_{i=1}^{C} -p_i * log_2(p_i)$$

$$S_0 = -\frac{9}{20}\log_2\frac{9}{20} - \frac{11}{20}\log_2\frac{11}{20} \approx 1$$



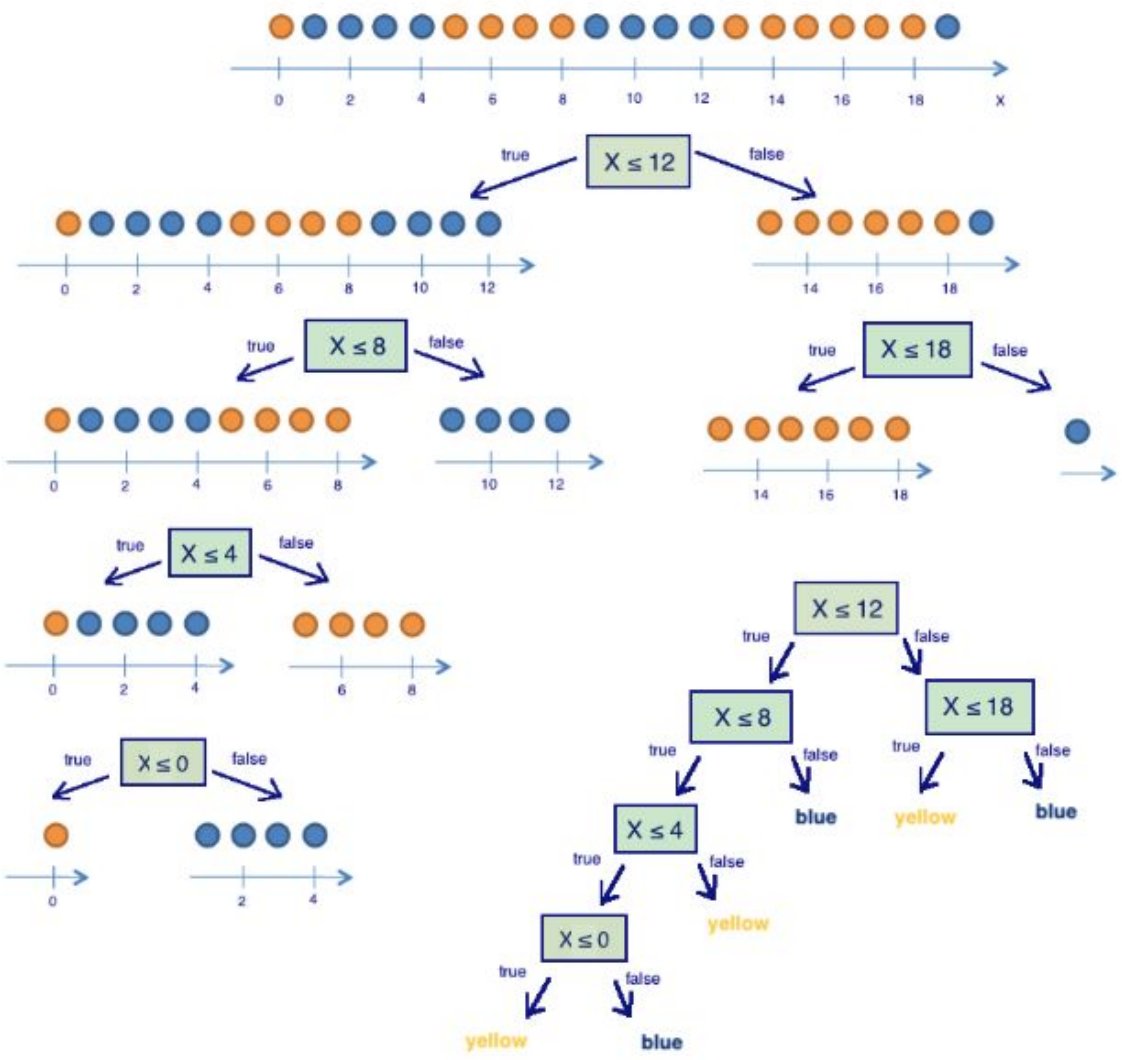
- True:

$$S_1 = -\frac{5}{13}\log_2\frac{5}{13} - \frac{8}{13}\log_2\frac{8}{13} \approx 0.96$$

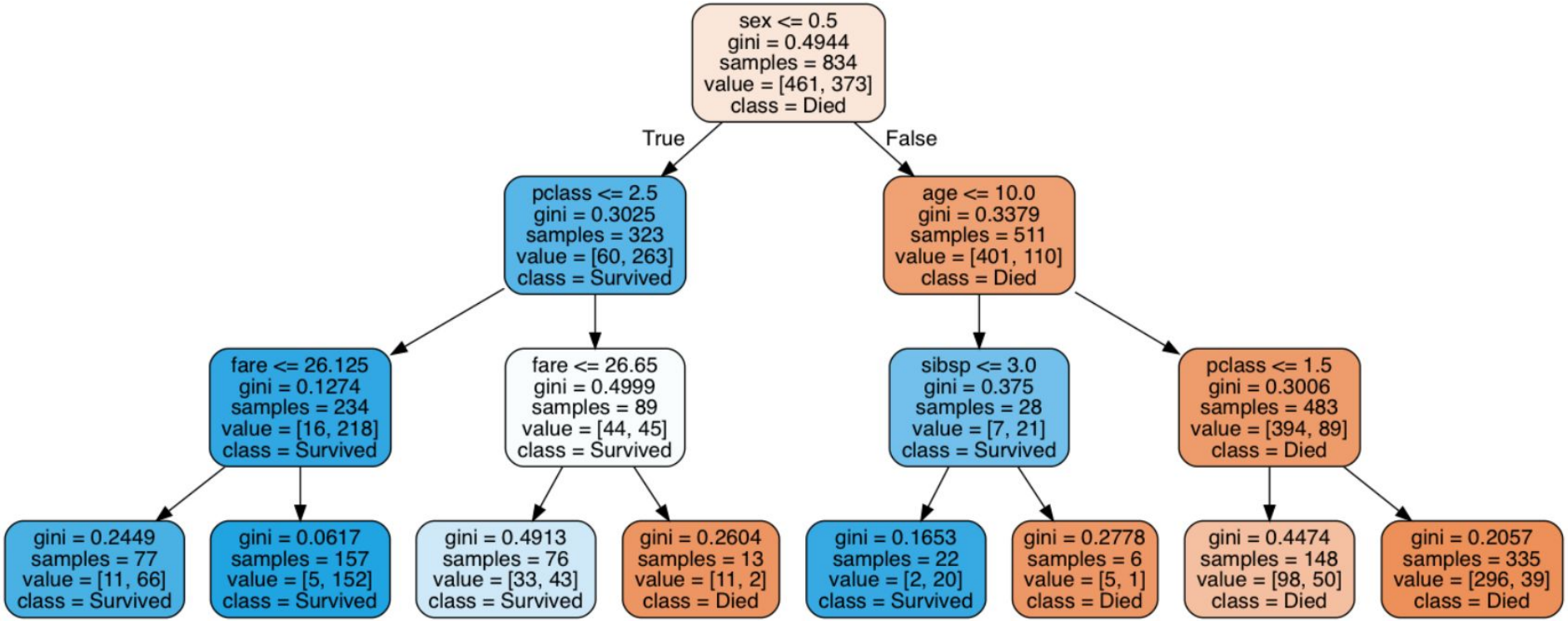
- False:

$$S_2 = -\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7} \approx 0.6$$

$$IG(Q) = S_O - \sum_{i=1}^{q}\frac{N_i}{N}S_i, \qquad IG(x \leq 12) = S_0 - \frac{13}{20}S_1 - \frac{7}{20}S_2 \approx 0.16.$$

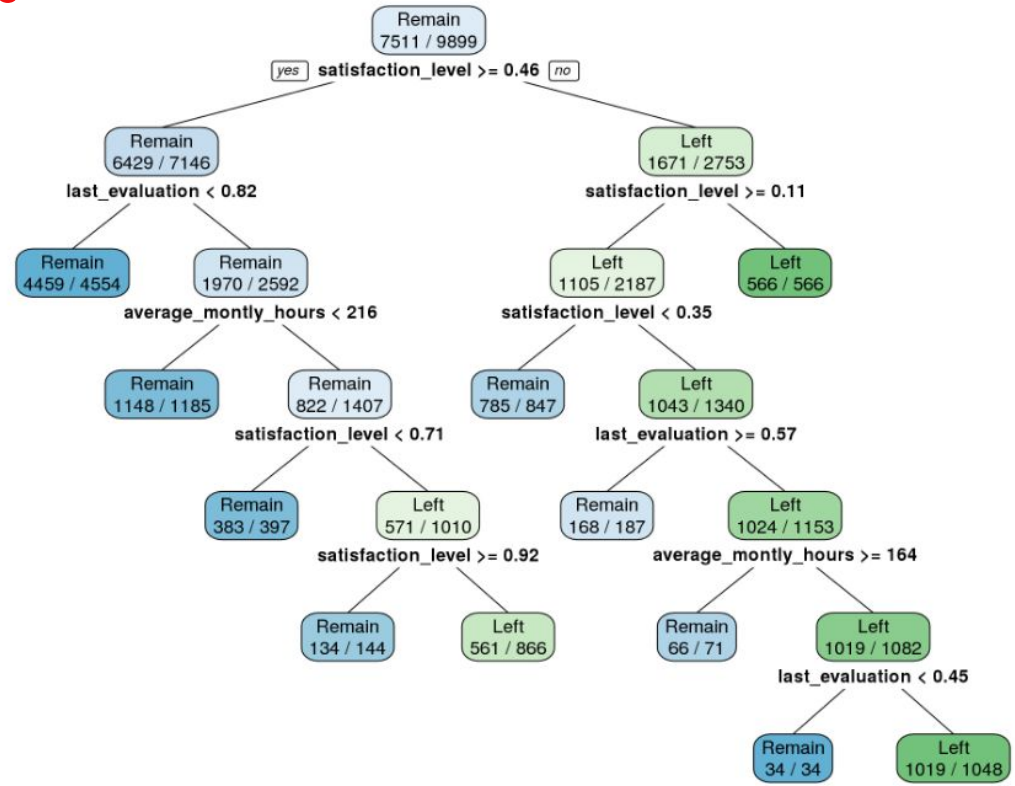

0 2 4 6 8 10 12 14 16 18 X
X ≤ 12
true
false
0 2 4 6 8 10 12
14 16 18
X ≤ 8
X ≤ 18
0 2 4 6 8
10 12
14 15 18
X ≤ 4
0 2 4
6 8
X ≤ 0
0
2 4
X ≤ 12
X ≤ 8
X ≤ 18
X ≤ 4
X ≤ 0
blue
yellow
blue
yellow
yellow
blue

Titanic - expanded



Rozhodovacie stromy - *pruning*

- Pre-pruning:
  - Zastavovacia podmienka
- Odstránenie nadbytočných uzlov - vyhodnocovanie chyby na validačných dátach
  - Minimum error
  - Smallest tree

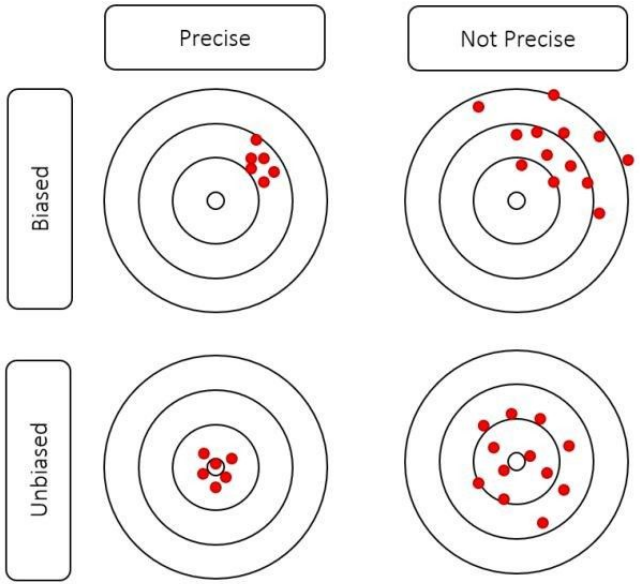


Chyby modelov (bias-variance tradeoff)

- Dajú sa rozdeliť do 3 častí:
  - Bias
    - Nakoľko sa líšia predpovede od pravdy
    - Underfitting
  - Variance
    - Nakoľko sa predpovede líšia navzájom
    - Overfitting
  - Irreducible error

$$error = (E[\widehat{f}(x)] - f(x))^2 + E[\widehat{f}(x)] - E[\widehat{f}(x)]] + \sigma_e^2$$

$$error = Bias^2 + Variance + IrreducibleError$$



Súborové učenie

- Rozhodovacie stromy - obvykle slabý klasifikátor
- Súborové učenie - spojením viacerých klasifikátorov získame lepší model
- Slabé klasifikátory sa môžu líšiť v:
  - Dátach
  - Hypotéza
  - Modeloch
  - Inicializácii
- Tri typy:
  - Boosting
  - Bagging
  - Stacking

Bagging - Bootstrap AGGregatING

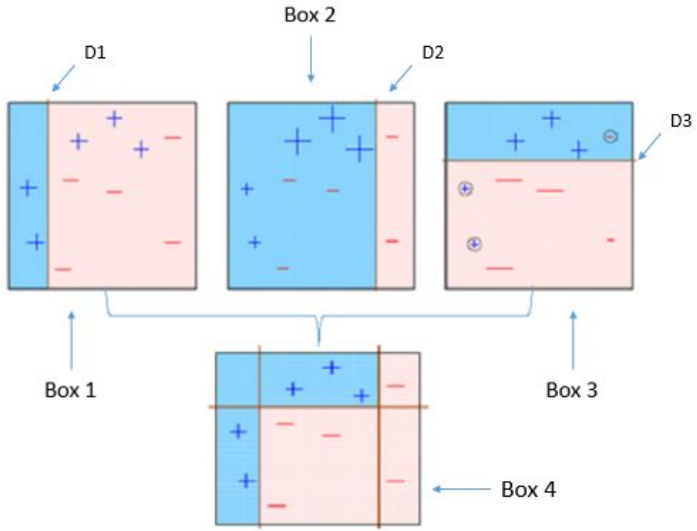
- Kombinuje viacero slabých modelov do jedného silného - často sú homogénne (rovnaký algoritmus ML)
- Každý model sa trénuje samostatne a spájajú sa:
  - Priemerovaním (regresia)
  - Hlasovaním (klasifikácia):
    - Najčastejšia trieda (hard-voting)
    - Priemer z pravdepodobností (soft-voting)
- Znižuje variance
- Reprezentatívnosť a nezávislosť

Bagging - napr. Random forest

- Zabezpečím nezávislosť:
  - Náhodne vyberám množinu pre každý strom (a môžu sa opakovať)
  - Náhodne sa pre stromy vyberá podmnožina vstupných príznakov
- Sklearn:
  - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
- Hyperparametre:
  - Počet stromov
  - Kriteriálna funkcia (gini/entrópia...)
  - Zastavovacia podmienka

Boosting

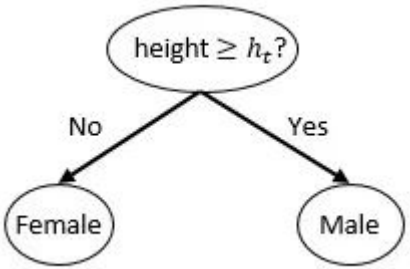
- Kombinuje viacero slabých modelov do jedného silného
- Modely sa trénujú jeden za druhým - snažia sa opraviť vzniknuté chyby
- Spájajú sa váhovaním alebo priemerovaním
- Znižuje bias (aj variance)



Zdroje obrázkov:
Schapire, R. E. and Freund, Y. (2012).Boosting: Foundations and Algorithms.MIT Press.

Boosting - napr. AdaBoost

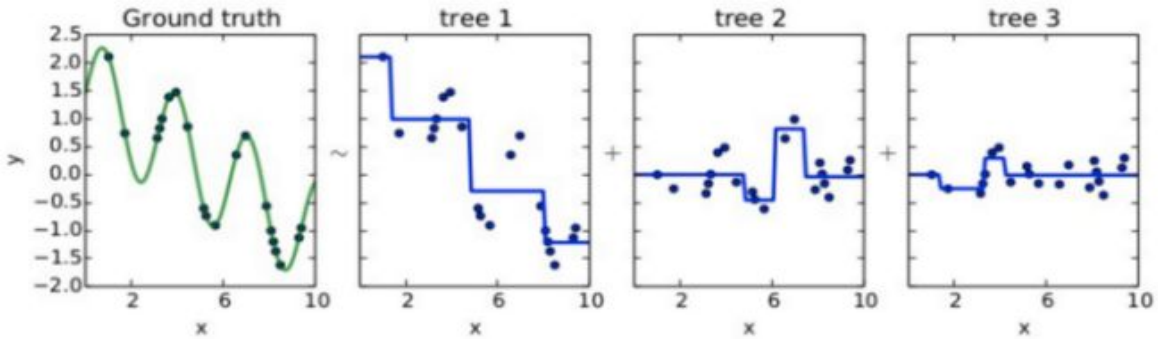
- Využíva *rozhodovacie pne* :) (rozhodovacie stromy hĺbky 1) ako slabé klasifikátory
- Prikladá váhy pre pne aj vzorky:
    1. Inicializujú sa váhy pre každú vzorku (rovná sa 1/počet vzoriek)
    2. Vytvorí sa rozhodovací pník
    3. Vypočíta sa chyba a váha daného pníka
    4. Prepočítajú sa váhy pre vzorky - ak boli klasifikované zle, ich váha sa zväčší
    5. Späť na krok 1 po zastavovaciu podmienku
    6. Konečné rozhodnutie - váhovaný priemer zo stromov



Zdroje obrázkov:
https://sefiks.com/2018/11/02/a-step-by-step-adaboost-example/

Boosting - napr. Gradient Boosting

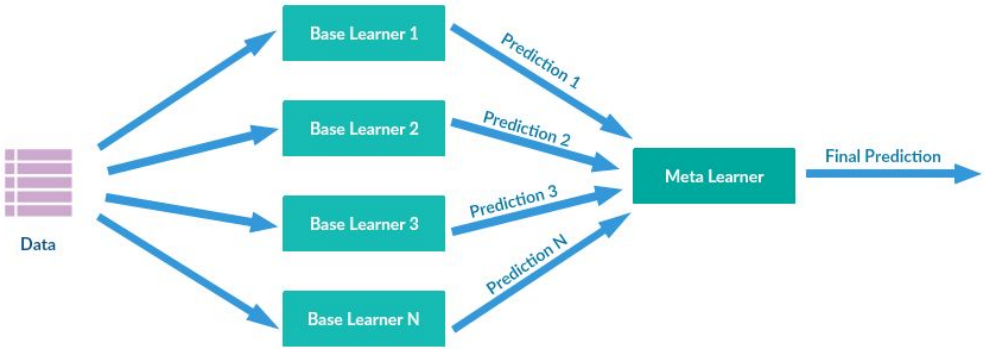
- Podobne ako AdaBoost, ale snažím sa opraviť reziduály
- Nové stromy sú pridávané aditívne



Zdroje obrázkov:
https://www.quora.com/How-would-you-explain-gradient-boosting-machine-learning-technique-in-no-more-than-300-words-to-non-science-major-college-students

Stacking

- Stacked generalization
- Spája silné klasifikátory
- Type modelov:
  - Level 0 - silné klasifikátory pre problém
  - Level 1 - model naučený spájať klasifikátory



Zdroje obrázkov:
https://mc.ai/icu-survival-prediction-using-ensemble-learning-stacking/