

PA153 Počítačové zpracování přirozeného jazyka

12 – Strojové učení a ZPJ

Jiří Materna, @JiriMaterna

Seznam.cz

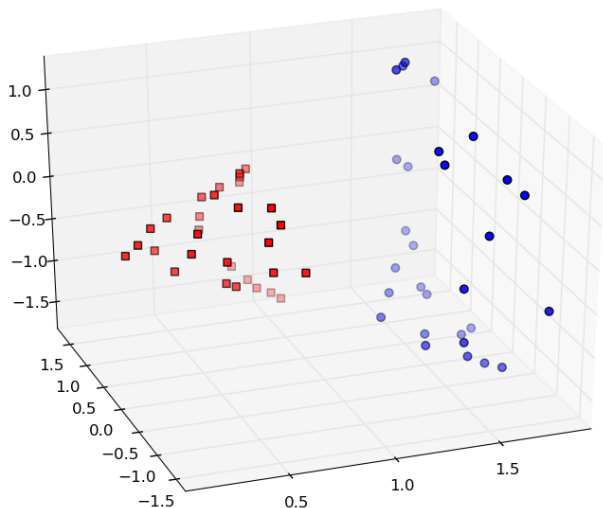
7. prosince 2015

- 1 Metody strojového učení
- 2 Klasifikace dokumentů
- 3 Modelování témat dokumentů
- 4 Jazykové modely

Strojové učení

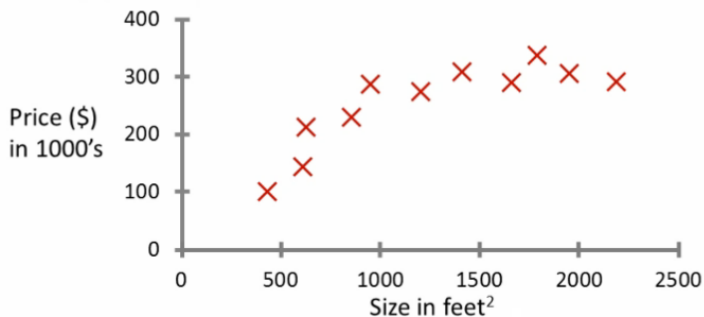
- učení s učitelem (supervised)
- učení bez učitele (unsupervised)
- kombinace předchozího (semi-supervised)
- zpětnovazební učení (reinforcement learning)
- optimalizační úloha

Klasifikační úloha



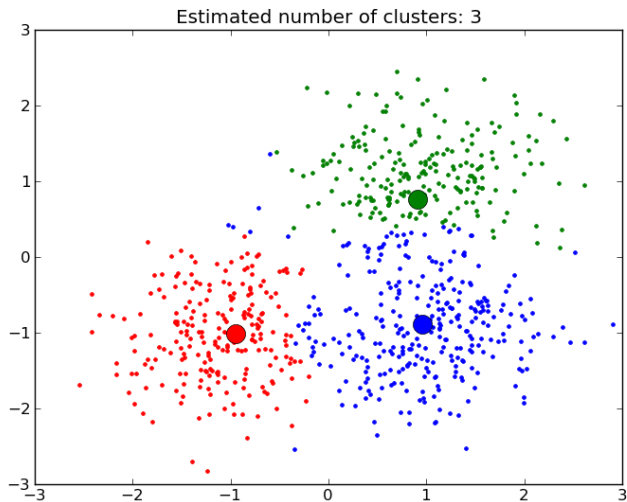
Regresní úloha

Housing price prediction.

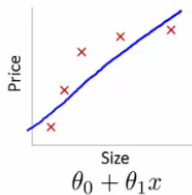


Zdroj: <https://class.coursera.org/ml/class>

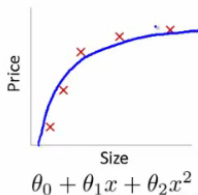
Shlukování



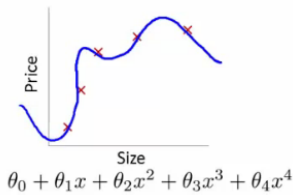
Problém s přeučováním



High bias
(underfit)



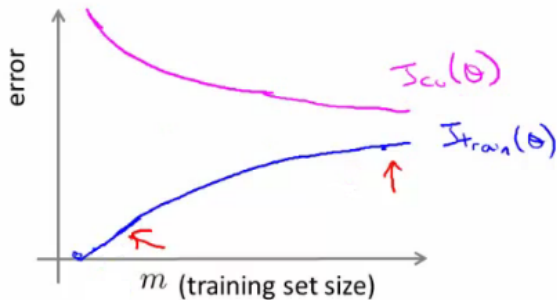
“Just right”



High variance
(overfit)

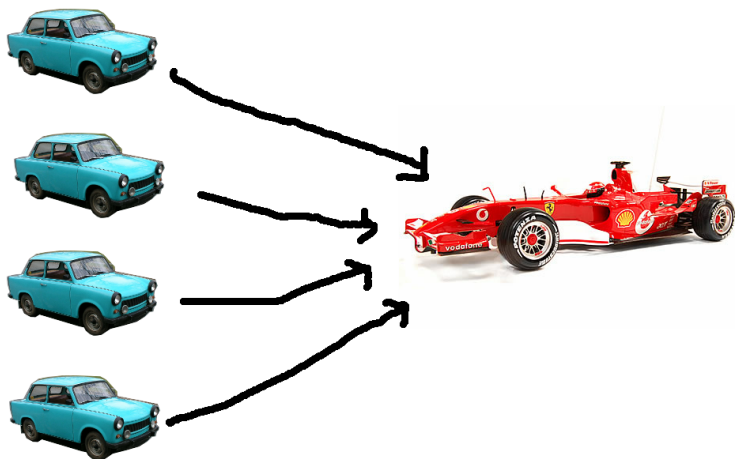
Zdroj: <https://class.coursera.org/ml/class>

Velikost trénovacích dat



Zdroj: <https://class.coursera.org/ml/class>

Boosting



Klasifikace dokumentů



Bag-of-words reprezentace dokumentů

- 1 the man walked the dog
- 2 the man took the dog to the park
- 3 the dog went to the park

[dog, man, park, the, to, took, walked, went]

- 1 [1, 1, 0, 1, 0, 0, 1, 0]
- 2 [1, 1, 1, 1, 1, 1, 0, 0]
- 3 [1, 0, 1, 1, 1, 0, 0, 1]

TF-IDF

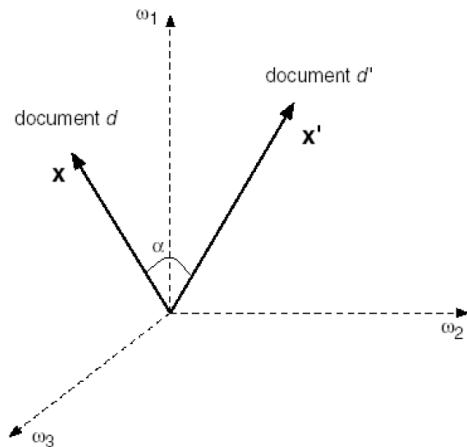
$$\text{TF}(t, d) = \frac{n_{t,d}}{\sum_k n_{k,d}} \quad (1)$$

$$\text{IDF}(t) = \log \frac{|D|}{|j : t_j \in d_j|} \quad (2)$$

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t) \quad (3)$$

- ① [0, 0.18, 0, 0, 0, 0, 0.48, 0]
- ② [0, 0.18, 0.18, 0, 0.18, 0.48, 0, 0]
- ③ [0, 0, 0.18, 0, 0.18, 0, 0, 0.48]

Modelování témat dokumentů

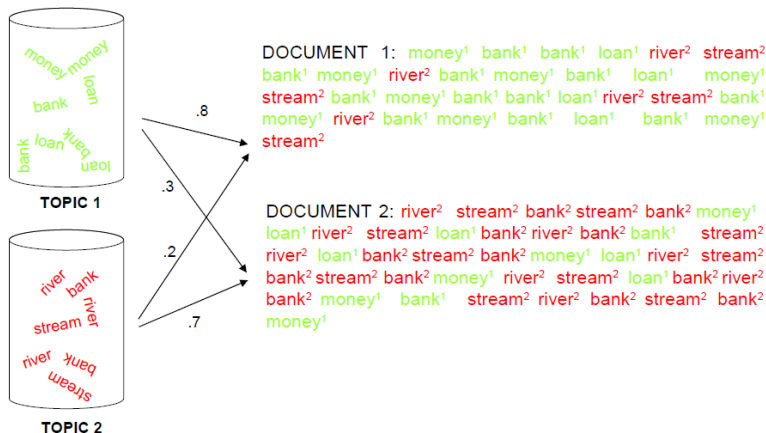


Latentní sémantická analýza

$$\begin{array}{ccccccc} & & X & & U & & \Sigma & & V^T \\ & & (\mathbf{d}_j) & & & & & & (\hat{\mathbf{d}}_j) \\ & & \downarrow & & & & & & \downarrow \\ (\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = & (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \left[\begin{array}{c} \mathbf{u}_1 \end{array} \right] \\ \dots \\ \left[\begin{array}{c} \mathbf{u}_l \end{array} \right] \end{bmatrix} & \cdot & \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix} & \cdot & \begin{bmatrix} \left[\begin{array}{c} \mathbf{v}_1 \end{array} \right] \\ \vdots \\ \left[\begin{array}{c} \mathbf{v}_l \end{array} \right] \end{bmatrix} \end{array}$$

Zdroj: http://en.wikipedia.org/wiki/Latent_semantic_analysis

Latentní Dirichletovská alokace



Zdroj: Probabilistic Topic Models, Mark Steyvers and Tom Griffiths, 2007

Jazykové modely

