Information Mining - winter semester 2019

Exercise sheet 8

Exercise 1: Linear regression with *R*

The example data[1] contain the number of new students in the last years.

(a) Create a scatter plot of the data

(b) Calculate the regression line and draw it into the plot

(c) The linear regression can be used as a prediction method. How many students will be expected for the year 2016?

(d) Does a correlation between the year and the number of new students exist? Make a correlation analysis according to Spearman.

Exercise 2: Instance based learning: k-nearest-neighbour

(a) Briefly sketch how a classification with k-nearest-neighbour is done.

(b) k-NN belongs to the so called lazylearning methods. What does this mean.

(c) When using k-NN a distance value is needed, which calculates the distance between the instances. Often the so called cosine similarity is used. The cosine similarity is defined as:

$$cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|},$$

where $\vec{a}$ and $\vec{b}$ are the two instances in vector space.

Given are the following four training instances:

$$\vec{a} = \begin{pmatrix} 0 \\ 3 \\ 5 \end{pmatrix}, \vec{b} = \begin{pmatrix} 3 \\ 3 \\ 8 \end{pmatrix}, \vec{c} = \begin{pmatrix} 2 \\ 6 \\ 1 \end{pmatrix}, \vec{d} = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix}$$

To which class would the instance $\vec{x} = (5, 8, 0)^T$ be assigned, if $\vec{a}$ and $\vec{b}$ belong to class X and $\vec{c}$ and $\vec{d}$ belong to class Y (2-NN-classification with cosine similarity as distance calculation should be used)?

Exercise 3: Information gain

Given the data in Table 1 we shall construct a decision tree. Let us assume the first split (root node) is made on the attribute *Taste*. What is the information gain associated with this attribute?

[1] `http://www.is.inf.uni-due.de/courses/im_ws19/uebung/data_r.csv`

Tabelle 1: Some data

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	Hot	Sour	Small
No	Hot	Salty	Large
Yes	Hot	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	Hot	Salty	Large

Exercise 4: Understanding questions

In which one of the following figures do you think the hypothesis has overfit the training set?

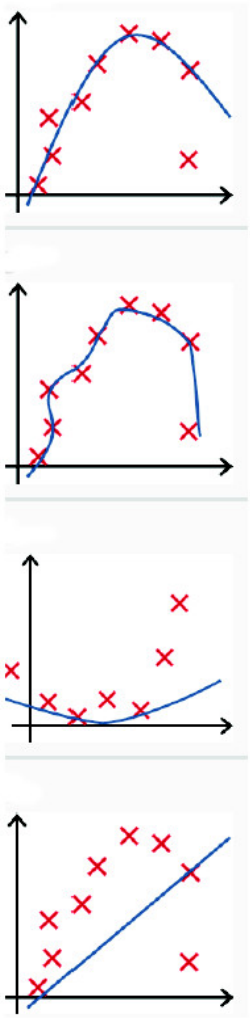

Abbildung 1: Overfitting examples. Pick the right one.