Information mining exercise, winter term 2022/23
Dr.-Ing. Benedikt Loepp (benedikt.loepp@uni-due.de)

Exercise sheet 4 Presentation on **24.11.2022**

Exercise 1: Overfitting

Explain briefly: What is overfitting? What are the causes and possible solutions?

Exercise 2: Linear regression

Given the following set of training data, with m examples:

x	y
3	4
2	1
4	3
0	1

Each row represents one training example. A linear model can be written as follows: $h(x) = w_0 + w_1 * x$. The cost function for learning the model parameters can be written as:

$$J(w_0, w_1) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^i) - y^i)^2 \quad (1)$$

- Suppose $w_0 = 1$ and $w_1 = 1.5$. Calculate $h(2)$.
- Calculate the cost for $w_0 = 0$ and $w_1 = 1$, i.e. $J(0, 1)$.

Exercise 3: Numeric prediction with *RapidMiner*

Now, we want to predict CPU speed with *RapidMiner*. The dataset is available in the Moodle course.

(a) Create a process with *RapidMiner* that learns a linear regression function.

(b) Provide the regression function.

(c) Make a prediction for the following data: `MYCT=270, MMIN=3000, MMAX=7000, CACH=120, CHMIN=12, CHMAX=32`

Exercise 4: Naïve Bayes

With the *Naïve Bayes* function the probability of H (hypothesis or event) given an E (evidence) is calculated:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Below, the weather data from the lecture are given. `play` is the class attribute (H), `day` is used as an identifier and is not part of the instances.

day	outlook	temperature	humidity	windy	play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

Figure 1: The weather data

Given are the following two instances:

- `outlook = rainy, temperature = hot, humidity = normal, windy = false`
- `outlook = overcast, temperature = hot, humidity = normal, windy = true`

(a) Calculate the probability $P(H|E)$ for both instances. If a problem with 0 frequency occurs (hint: this happens for the `outlook` attribute), use modified probability estimates (as seen in the lecture, with the help of the Laplace estimator and $\mu = 1$).

(b) Explain to which class the instances will be sorted to.