Information Mining - winter semester 2020/21

Exercise sheet 4

Excercise 1: Understanding

What is overfitting? What are the causes and possible solutions?

learner performs too good in training set but bad in test set.
model is too complex.

1.simplify the model , e.g. pruning , drop out ...
2.regularization , to weaken noises.

Excercise 2: Linear Regression

We will use the following training set of a small sample of different students' performances:

x	y
3	4
2	1
4	3
0	1

Each row is one training example. In the lecture a linear model was introduced as $h(x) = w_0 + w_1 * x$. We shall use m to denote the number of training examples. Also note in the lecture we introduced the notion of cost function for learning the model parameters. Such a cost function can be written as:

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

- What is the cost for $w_0 = 0$ and $w_1 = 1$, i.e. for J(0,1)? 0.5
- Suppose $w_0 = 1$ and $w_1 = 1.5$. What is h(2)? 4

Excercise 3: Numeric prediction with *RapidMiner*

The speed of a CPU[1] should be predicted.

(a) Create a process with *RapidMiner*, which should learn a linear regression function.

(b) How does the regression function look like?

(c) Which value will be predicted for the following data?

```
MYCT=270, MMIN=3000, MMAX=7000, CACH=120, CHMIN=12, CHMAX=32
```

[1] dataA10.arff

Excercise 4: Naive Bayes

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

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

The weather data from the lecture are given. `play` is the class attribute (event), `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 $Pr(H|E)$ for both instances.

Use the modified probability estimates (as seen in the lecture using the Laplace estimation) with $\mu = 1$ (Attribute `outlook`), if a problem with 0 frequency occurs.

(b) To which class will the instances be sorted to?