8. Exercise Sheet **Summer Term 2023**
Statistical Methods for Data Analyses A **Prof. W. Rhode**
Submission: 13.06.2022 23:59 **Dr. M. Linhoff**

Time	Group	Submission in Moodle; Mails with subject: [SMD2023]
Th. 12:00–13:00	A	lukas.beiske@udo.edu and tristan.gradetzke@udo.edu
Fr. 08:45–09:45	B	jonas.hackfeld@ruhr-uni-bochum.de and ludwig.neste@udo.edu
Fr. 10:00–11:00	C	stefan.froese@udo.edu and vincent.latko@udo.edu

Exercise 16 *Naive Bayes: Soccer* **5 p.**

Bayes' theorem states:

$$P(S|W) = \frac{P(W|S) \cdot P(S)}{P(W)} \tag{1}$$

(a) Prove Bayes' theorem (1) using the definition of conditional probability.

In this task S describes whether soccer is played or not. W describes the weather condition, which is described by four attributes. The data set in table 1 is available to you.

Attributes	S = yes	S = no
Wind	low(6), high(3)	low(2), high(3)
Humidity	high(3), normal(6)	high(4), normal(1)
Temperature	hot(0), mild(6), cold(3)	hot(1), mild(1), cold(3)
Forecast	sunny(2), cloudy(4), rainy(3)	sunny(1), cloudy(1), rainy(3)

Table 1: Data set. (The brackets indicate how often the corresponding value was measured).

b) Weather conditions for today can be found in table 2. What is the probability that soccer will be played today?

Attribute	Value
Wind	high
Humidity	high
Temperature	cold
Forecast	sunny

Table 2: Weather conditions today.

Hints:

1. You can use (2) under the naive assumption that the attributes x_i are independent:

$$P(W|S) = \prod_i P(x_i|S) \tag{2}$$

2. Consider what the normalization $P(W)$ is composed of.

c) Suppose you should now calculate what is the probability of playing soccer tomorrow (weather conditions see table 3). What problem occurs and how can you solve it?

Attribute	Value
Wind	low
Humidity	high
Temperature	hot
Forecast	sunny

Table 3: Weather conditions tomorrow.

Exercise 17 *Binary Decission Tree: The First Decission* **5 p.**

The file `soccer.csv` provides the dataset in table 4. The attributes are

- `temperature`: Temperature in degree Celsius.
- `weather_forecast`: Sunny, rainy or cloudy; the overall weather condition.
- `humidity`: Humidity in percent.
- `wind`: Statement whether it is windy right now.
- `soccer`: Statement whether it is worth going to play soccer.

The last attribute is your decision target. In this task, you have to find the first cut of a *binary* decision tree for this purpose.

(a) Calculate (by hand) the entropy of the tree's root.

(b) Calculate (by hand) the information gain if a cut is made on the attribute `wind`.

(c) For the remaining attributes, implement and plot the information gain as a function of different cuts. Distinguish between ordinal, nominal, and cardinal attributes. Treat (in particular) the weather forecast as a nominal attribute to produce a unified solution. Consider how to implement cuts on the different attribute classes.

(d) Which attribute is suited best to derive a decision?

Table 4: Dataset: "Should I play soccer?"

`temperature` / °C	`weather_forecast`	`humidity` / %	`wind`	`soccer`
29.4	sunny	85	False	False
26.7	sunny	90	True	False
28.3	cloudy	78	False	True
21.1	rainy	96	False	True
20.0	rainy	80	False	True
18.3	rainy	70	True	False
17.8	cloudy	65	True	True
22.2	sunny	95	False	False
20.6	sunny	70	False	True
23.9	rainy	80	False	True
23.9	sunny	70	True	True
22.2	cloudy	90	True	True
27.2	cloudy	75	False	True
21.7	rainy	80	True	False