Information Mining - winter semester 2019

Exercise sheet 7

Exercise 1: Understanding: Evaluation metrics

(a) While classifying the following four outcomes can occur. Please shortly explain the meaning of these terms and give an example.

- true positive, TP
- true negative, TN
- false negative, FN
- false positive, FP

(b) How are the following measures defined?

- accuracy
- mean-squared error
- precision and recall
- F1-measure

Exercise 2: Deeper Understanding

Imagine, you are working with "Zeit.de" and you want to develop a machine learning algorithm which predicts the number of views on the articles.

Your analysis is based on features like author name, number of articles written by the same author on Zeit.de in past and a few other features. Which of the following evaluation metric would you choose in that case?

(a) Mean Square Error

(b) Accuracy

(c) F1 Score

Exercise 3: R and RStudio

R[1] is a tool for statistic analysis and graph-generation. A GUI for *R* is *RStudio*.[2] Install *R* and *RStudio* and get to know the Application. A installation guide and a help are available on the given websites.

Exercise 4: Statistics with *R*

Calculate the following values using the example data[3]:

[1] `http://www.r-project.org`

[2] `http://www.rstudio.com/products/RStudio/#Desktop`

[3] `http://www.is.inf.uni-due.de/courses/im_ws19/uebung/data_a16.csv`

(a) Median of the age

(b) Arithmetic mean of the earnings

(c) Number of married people

Draw a histogram with the data from the age column. The histogram should contain 6 intervals.

Exercise 5: Significance test and correlation with*R*

Use the example data from exercise 2.

(a) Is there a significant difference between the earning of people who responded and those who did not respond? Use the t-Test to calculate.

(b) Does a correlation between the age and the earning exist? Make a correlation analysis according to Spearman.