Exercise for Machine Learning (SS 20)

Assignment 2: Naive Bayes and Text Classification

Prof. Dr. Steffen Staab, steffen.staab@ipvs.uni-stuttgart.de

Alex Baier, alex.baier@ipvs.uni-stuttgart.de

Janik Hager, janik-manel.hager@ipvs.uni-stuttgart.de

Ramin Hedeshy, ramin.hedeshy@ipvs.uni-stuttgart.de

Analytic Computing, IPVS, University of Stuttgart

Submit your solution in Ilias as either PDF for theory assignments or Jupyter notebook for practical assignments.
Mention the names of all group members and their immatriculation numbers in the file.
Submission is possible until the following Monday, 11.05.2020, at 14:00.

1 Simple Bayes

1. Box 1 contains 8 apples and 4 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability. What is the probability of choosing an apple? If an apple is chosen, what is the probability that it came from box 1?

2. The blue M&M was introduced in 1995. Before then, the color mix in a bag of plain M&Ms was: 30% Brown, 20% Yellow, 20% Red, 10% Green, 10% Orange, 10% Tan. Afterward it was: 24% Blue , 20% Green, 16% Orange, 14% Yellow, 13% Red, 13% Brown.

   A friend of mine has two bags of M&Ms, and he tells me that one is from 1994 and one from 1996. He won't tell me which is which, but he gives me one M&M from each bag. One is yellow and one is green. What is the probability that the yellow M&M came from the 1994 bag?

2 Spam Classification with Naive Bayes

Please download the Jupyter notebook *assignment2.ipynb* and the dataset *spamham.txt*. Follow the instructions in the Jupyter notebook.

3 kNN for Text Classification

Research and discuss how you could use a k-nearest neighbor classifier for text classification. You should at least answer these questions:

- How do you represent the text?

- What distance function do you use?
- What decision rule do you use?

Provide an example for your representation of the text and how your classification decision is made based on the distance function and decision rule. Explain the advantages and disadvantages of your approach.

4 kNN in High-Dimensional Feature Spaces

For all students other than B.Sc. Data Science:
Research and discuss why kNN might fail for high dimensional feature spaces. Identify and explain one approach for solving or circumventing this problem.