University of Stuttgart

ANALYTIC COMPUTING

Machine Learning Exercise (SS 21)

Assignment 2: Naive Bayes and Text Classification

Dr. Decky Aspandi
decky.aspandi-latif@ipvs.uni-stuttgart.de

Akram Hosseini
Akram.Hosseini@ipvs.uni-stuttgart.de

Daniel Frank
daniel.frank@ipvs.uni-stuttgart.de

This assignment sheet consists of 5 pages with the following 4 tasks:

Submit your solution in ILIAS as a single PDF file.[1] Make sure to list your full name and immatriculation number at the start of the file. Optionally, you can *additionally* upload source files (e.g. PPTX files). Remember to fill out the exercise slot and exercise presentation polls linked in ILIAS. If you have any questions, feel free to ask them in the excercise forum in ILIAS.

Submission is open until Tuesday, 11 May 2021, 23:59 PM.

[1] Your drawing software probably allows to export as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

Task 1: Simple Bayes (20 Points)

1. **Task (10 Points)**: Box 1 contains 4 apples and 10 oranges. Box 2 contains 6 apples and 8 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. **Task (10 Points)**: The blue M&M was introduced in 1995. Before then, the color mix in a bag of plain M&Ms was: 20% Brown, 30% Yellow, 10% Red, 20% Green, 10% Orange, 10% Tan. Afterward it was: 20% Blue , 24% Green, 14% Orange, 16% Yellow, 13% Red, 13% Brown.

   A friend of mine has two bags of M&Ms, and she tells me that one is from 1994 and one from 1996. She won't tell me which is which, but she 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?

Task 2: Fake News Classification with Naive Bayes (20 Points)

1. **Task (20 Points)**: Please download the Jupyter notebook *assignment2.ipynb* and the dataset *liar.txt*. Follow the instructions in the Jupyter notebook.

Task 3: kNN for Text Classification (30 Points)

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

1. **Task (10 Points)**: How do you represent the text?
2. **Task (10 Points)**: What distance function do you use?
3. **Task (10 Points)**: 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.

Task 4: kNN in High-Dimensional Feature Spaces (30 Points)

1. **Task (10 Points)**: Research and discuss why kNN might fail for high dimensional feature spaces.
2. **Task (20 Points)**: Identify and explain one approach for solving or circumventing this problem.