Exercise 2 - Decision Theory and Density Estimation

Winter term 2019/2020

Important

Please solve the assignments in groups of 3 to 4 students. The solutions are going to be presented and discussed after the submission deadline. Sample solutions will not be uploaded. However, you are free to share correct solutions with your colleagues **after they have been graded**. Please submit your solutions via Moodle **and** in printed form. Only one member of the group has to submit the solutions. Therefore, make sure to specify the names of all group members. Please do not submit hand-written solutions, rather use proper type-setting software like LaTeX or other comparable programs.

Your homework will be corrected and given back to you. Correct solutions are rewarded with a bonus for the exam (max. 10 percent, if all solutions submitted are correct). **Please note:** You have to pass the exam **without the bonus points**! *(i.e. it is not possible to turn 5.0 into 4.0)* The solutions have to be your own work. If you plagiarize, you will lose all bonus points!

Further remarks:

- Code assignments have to be done in `Python`
- The following packages are allowed: `numpy`, `pandas`
  (please ask, if you want to use a specific package not mentioned here)
- **Do not use already implemented models** (e. g. from `scikit-learn`)

1 Bayesian Decision Theory

a) Bayes' Rule (1 point)

State Bayes' rule and state the name of each term in the equation.

Solution:

b) Decision Boundary (1 point)

Which condition holds at the optimal decision boundary? In a binary classification problem, when do we prefer class A over class B? Why is it not necessary to normalize the probabilities on both sides of the inequality?

Solution:

c) Naïve Bayes (5 points)

You are planning a nice trip to the forest to collect some delicious mushrooms. However, you are not an expert for mushrooms and afraid of picking poisonous ones. However, you have a data set called `mushrooms.csv` describing the shape, color and habitat of different mushrooms and whether they are edible (type e) or poisonous (type p). Implement a binary naïve Bayes classifier using `Python` and `numpy` / `pandas` and train it on the mushrooms data set. Use 10 % of the data set as test set and report your accuracy on this test set.

Solution:

d) Bonus Question (1 point)

You work for a machine learning startup which specializes in text classification for automatic scam detection in social networks. You are required by law to explain in detail why your system did not filter out content which was scam or why it did filter out normal content. Given that both models are suitable, would you prefer a naïve Bayes model or a deep neural network? Why?

Solution:

2 Density Estimation

a) Non-Parametric Density Estimation (3 points)

You want to estimate the density for the data stored in the file `density_data_train.csv`. Implement either a kernel density estimator using a Gaussian kernel trying different values for h **or** a k-nearest neighbors estimator trying different values for k. Which value for h or k works best? Plot the densities in a suitable interval.

Solution: