# Exercise 2 - Decision Theory and Density Estimation

August 16, 2019

# 1 Bayesian Decision Theory

a) Bayes' Rule (1 point)

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

b) Decision Boundary (1 point)

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

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. You got a dataset *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 and train it on the mushrooms dataset. Use 10% of the dataset as test set and report your accuracy on this test set.

d) Bonus Question (1 point)

You are working 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 a 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?

## 2 Density Estimation

a) Non-Parametric Density Estimation (3 points)

You want to estimate the density for the data in the file *density_data_train.csv*. Implement either a kernel density estimator using a Gaussian kernel and try different values of $\sigma$ or a k-nearest neighbors model and try different numbers of neighbors $k$. Which value for $\sigma$ or $k$ works best? Plot the densities on the test data (*density_data_test.csv*) in the interval $[-8, 8]$.