Exercises for
Pattern Recognition
Sebastian Käppler, Nooshin Haji Ghassemi
Assignment 1, 26/27.10.2014

## General Information:

| | |
|---|---|
| Exercises (1 SWS): | Mo 12:15 – 13:30 (0.151-115) and Tue 08:30 – 09:45 (0.151-115) |
| Certificate: | Oral exam at the end of the semester |
| Contact: | sebastian.kaeppler@fau.de, nooshin.haji@fau.de |

# $k$-NN and Bayesian classification

**Exercise 1** Create a k-Nearest Neighbor classifier for the Classification Toolbox, where k is a user-defined variable. Compare its output to the output of the already existing Nearest Neighbor classifier.

Hint: The k-Nearest Neighbor (k-NN) classifier calculates the distance of a new feature to the features of the training set. The new feature is assigned to the class to which the majority of the k nearest neighbors belongs to.

**Exercise 2** A computer science student is annoyed that the two thirds of the e-mails he/she receives is spam. Therefore, he/she decides to write a classifier that should decide whether an incoming e-mail is spam (class y=1) or ham (class y=0). For classification the Bayes classifier is used. The student notices that in spam and ham mails, certain words occur with different probability. Therefore, the student bases the classification on the words $\boldsymbol{x} = \{Viagra, bet, student, sports, cinema\}$. By inspecting all his/her previous mails, the student estimates that in the ham mails, the probabilities are 0% for Viagra, 10% for bet, 40% for student, 30% for sports, and 10% for cinema. In the spam mails, the words Viagra occurs in 50%, bet in 30%, student in 5%, and sports and cinema in 2% of the mails.

The student does not count how often each word occurs, but only whether it is present in the mail. For simplicity, he/she assumes that the words occur independently.

(a) Has the student considered all the postulates of pattern recognition?

(b) Write down the priors for an e-mail being spam or ham.

(c) Write down the class-conditional probabilities $p(\boldsymbol{x}|y = 0)$ and $p(\boldsymbol{x}|y = 1)$ for an arbitrary feature vector $\boldsymbol{x}$. Hint: Bernoulli distributions for each feature $p(x_i|y = 0) = p(x_i = 1|y = 0)^{x_i} \cdot (1 - p(x_i = 1|y = 0))^{1-x_i}$

(d) Write down the Bayesian decision rule for the spam classification problem. Classify the following e-mail using the decision rule:
*Hi, As we talked about yesterday, I want to make a bet with you about the upcoming soccer match. I clearly know more about sports than you. I bet 5$ against Nürnberg.*