Exercise 11: Spam Classification with Naïve Bayes

Lecture Information Processing and Communication

Jörn Anemüller, June 2022

Submit solutions until Tuesday 2022-07-05, 23:59h, by uploading to your group's exercise folder on cs.uol.de. You may submit your solutions in groups of at most two students.

1. Spam Classification with Naïve Bayes

Implement the Naïve Bayes classification algorithm for the email-spam detection task and apply it to a small example dataset.

A spam/no-spam dataset is available at the UC-Irvine machine learning repository that we already used earlier under the address

`http://archive.ics.uci.edu/ml/machine-learning-databases/spambase/`

Note that there is also a documentation file that describes the data format used. The provided script (`script_spam.m`) loads the data, converts them into an appropriate condensed format according to the conventions used earlier, and subdivides them into training ($\sim$80%) and test data ($\sim$20%).

Implement the algorithm from the lecture in matlab or python. Which accuracy of correctly classified spam/no-spam emails does the algorithm achieve on training and test data, respectively?