Technische Universität München
Fakultät für Informatik
Prof. N. Navab, Ph.D.

SS 2016
Exercise Sheet Praktikum Machine Learning
May 9th, 2016

Exercises in Praktikum Machine Learning

The code available online performs classification with random forests. All the information you need is contained in the script *test_classification_forests.m*, which runs without any modification the random forest classifier on a toy example in 2D. You can have a look at the other files for deeper understanding, but there is no reason to change anything in them in the context of this exercise.

The exercise consists in studying experimentally the impact of some parameters of the random forest classifier when other parameters are fixed. Two synthetic datasets are proposed: they are automatically generated in the code (just uncomment the corresponding portion of code). The first dataset proposes to classify points sampled from two 2D gaussian distributions. The second proposes to perform binary classification in a high-dimensional feature space, where for simplification feature values are binary, i.e. a split is entirely determined by the choice of a dimension (threshold automatically set to 0.5).

You can directly refer to the comments included in the code for more information.

a) **2D Gaussian data**

(i) Study the performance (i.e. the error rate) of 1 tree on both the training set and the test set when the depth increases (depth varying between 2 and 20). What do you observe?

(ii) Study the performance of the random forest classifier when the number of tree increases.

b) (**High dimensional data**)

(i) Study the performance of a random forest when the number of tries varies, from 1 (completely random trees) to d (trees all correlated), where d is the dimension of the feature space.

(ii) By playing with both the number of relevant features in this high-dimensional dataset and the number of tries, study the robustness of random forests to irrelevant dimensions.