Intelligent Data Analysis

Tutorial 4
Random forest

Paul Prasse
Ahmed Abdelwahab
Dr. Niels Landwehr
Prof. Tobias Scheffer

Date: 15-05-2015

Goals

In this tutorial we will use the previous tutorial (decision trees) to build random forest. You should implement a simplified version of random forest learning method and apply this algorithm to the testing data. The goal is a deeper understanding of this learning process, the functioning of random forest and decision trees and their advantages and disadvantages.

Problem Setting

We will use the same data to compare between decision trees and random forest. As a recall, the dataset contains computed tomography images of 80 heart attack patients (see `spect.mat`). Each of these images is marked as conspicuous ($y_i = 1$) or inconspicious ($y_i = 0$). The data consists of 22 binary attributes, ie 22 different investigations. The goal is to make a prediction about the label for an additional set of 187 tomography images $Xtest$.

Task 1

Write the following MATLAB function

```
function Forest = learnForest(X,y,n)
```

which trains n decision trees on a sampled dataset from the training data X and the corresponding classes y. Each tree is trained on its own sampled dataset (with replacement). Feel free to use the functions implemented on the previous tutorial (learnTree with k=1).

Task 2

Write the following MATLAB function

```
function xClass = classForest(Forest,X)
```

which returns the classes of the dataset X using the random forest $Forest$. A class is predicted for a instance if the majority of the trees in the random forest agreed on.

Task 3

Train the forest using dataset Xtr and their classes Ytr. Compute the accuracy of the random forest classification on the dataset $Xtest$ by comparing it to $Ytest$. Then, compare the accuracy of the random forest and the decision trees. Where is the difference comes from?

Worth mentioning

Two points were simplified in the implemented Random forest.

- After selecting each bagging sample, Random forest randomly select a set of attributes before training the decision tree. Here for simplicity we trained the trees on all the attributes.
- Random forest is evaluated using out of bag error. In which, each instance is classified using all the trees in which the instance wasn't a part of their training set.