Introduction to Learning and Intelligent Systems - Spring 2015

Martin Ivanov (ivanovma@student.ethz.ch)
Can Tuerk (can.tuerk@juniors.ethz.ch)
Jens Hauser(jhauser@student.ethz.ch)

March 28, 2015

Project 2 : Two-Label Classification

Problem description

In this machine learning setting we received datafiles of biomedical images of human tissue which has to be classified into two categories - one with seven types and one with three. The data is beside the two classes, structured into nine variables representing parameters about geometrical and texture-related features and two variables which consist of four and forty binary columns in a one-of-k format.

Unfortunately the training data does not represent all possible combinations of the different subtypes of the two classes. Also the weights of the data differ enormously. We observed the following classes and weights within the training data:

{class(1,0)}: 3634 - {class(1,1)}: 1644 - {class(2,0)}: 308 - {class(2,1)}: 6827 - {class(3,1)}: 861 - {class(4,1)}: 42 - {class(4,2)}: 28 - {class(5,2)}: 239 - {class(6,1)}: 385 - {class(7,2)}: 546

Solution

Our solution with a best peformance of about 0.14 on the validation data due to the given loss function is mainly based on a two stage decision tree approach.

For our best result we used the randomized decision tree classifier 'ExtraTreesClassifier' of the scikit-learn package. First we trained our classifier on the first category of our training data and made predictions based on the best estimator found by cross validation and grid search on the hyper parameters and hereafter we did run the same classification procedure for the second category. The selected features and also the hyper parameters for this approach are provided in the attached code file.

Different Approaches

Before getting the above result, we carried out the following actions for improvment of our earlier results. On all of the following approaches we spend a lot of time to identify the best subset of our features. Unfortunately we runned out of time to get such a result as provided in the harder baseline.

- Classification based on random forest
- Classification on the principal components of the data
- K-nearest neighbour classification
- Naive Bayes classification
- Logistic Regression
- Stochastic Gradient Descendent classification
- Support Vector classification
- Unsupervised Learning tasks on the validation data to identify features