COL774 : Assignment 4 (Part II)

Aditi Singla
2014CS50277

28th April 2017

Note : All the training and testing done has been recorded in the file "report" and has been submitted along with the codes and this report.

1 Evaluation on basic algorithms:

1. Using SVM :

   - Linear SVM : From the file attributes.txt, it was noted that the main features were only 231 and so not all the features were necessary to train the SVM model. So PCA was applied twice on the training data to get the following number of projections and corresponding accuracies :
     - 20 projections : 80.311% (Training data: around 20 secs) & 76.235% (Cross-Validation: around 58 secs)
     - 250 projections : 60.659% (Training data: around 4 mins) & 76.215% (Cross-Validation: around 10 mins)
   - Gaussian SVM : Again, PCA was applied on the training data to get 250 projections. Since there were 231 important features, 250 seemed a reasonable choice. Accuracies observed were as follows :
     - 250 projections : 99.226% (Training data: around 130 mins) & 69.495% (Cross-Validation: around 670 mins i.e around 11 hours)

2. Using Decision Trees and Random Forests:

   - Decision Trees : This algorithm was run by varying min_samples_split, min_samples_leaf, max_depth features and the training accuracy was around 88.7% both the times, and test accuracy was around 94.3%.
   - Random Forests : This algorithm was run varying the n_estimators (number of trees), max_features, and bootstrapping. Accuracies on training data were 99-100% and cross-validation and test accuracies were 93-94%. (Details in "report" attached).

3. Using Naive Bayes:

   - BernoulliNB() : This algorithm was run on the complete training data and obtained around 81.5% accuracy on the training data.
   - GaussianNB() : This algorithm was again run on the complete training data and obtained around 77.59% accuracy on the training data.

On cross-validation, the accuracies were expected to reduce and hence this was not pursued much further.

The best accuracies were obtained for random forest classifier on varying different parameters, hence I chose to pursue it further in the second part for the competition.

2 Competition on Kaggle

Few features were varied in various ranges which were n_estimators (number of trees in the forests), max_features (maximum number of features that must be used), bootstrapping (replacement policy), max_depth (Maximum depth that can be reached in any tree), random_state (Seed to make it deterministic) & criterion (Entropy or Gini).

Accuracy increased with the number of estimators till values around 200-300 and then reduced. For max_features, good values were obtained at around 20-30% of the data (value = 0.2-0.3). Practically, accuracy should be better for bootstrapping set as true, but by observation, it was better for false. Entropy criterion gave better results. Seeding was done to get the results deterministically so that the results can be reproduced.