MSBD 6000B Deep Learning

Project 1

Student Name: Jingyi Wang
Student ID: 20469513
Due date: 11:59pm, Nov 16, 2017

1. Overview

This project investigates methods of handling specific classification problem with basic machine learning models. The training set includes 3220 instances with corresponding class labels, and each of them has in total 57 continuous features. By constructing various classifiers and comparing their quality on the training set with cross validation, a final model with the best performance will be induced to finish the task of classification on instances in testing set with 1380 instances.

2. Experiment

Basically, experiment will be conducted following the general procedures of binary classification. Since the dataset has good balance between positive and negative instances, stratified sampling may not be necessary. In order to avoid overfitting, 10-fold cross validation is performed for each classifier. A final model will be selected based on the average accuracy of predictions on both training and validation sets.

2.0 Preprocessing

One observation of the original training data is that mean values of each attribute vary from 0.005 to 278.6 and standard deviation vary from 0.07 to 523.6. Hence, normalization can stable the data and might make it easy for classifiers to fit the data later on. As an aside, the dataset is complete in the sense that it has no value in any attribute or label. Classifiers can use all the 57 continuous attributes to complete the task, though, the procedure of feature selections might make a difference for those classifiers that tend to overfit the data.

2.1 Classification

Before experiment on several candidate models, Naive Bayesian classifier is selected to be the baseline. Other classifiers under experiment involve logistic regression model, linear support vector classifier, SVM with different kernels and tuned parameters, Decision Tree, and ensemble methods including Adaboost with base model of simplified decision tree and random forest classifiers.

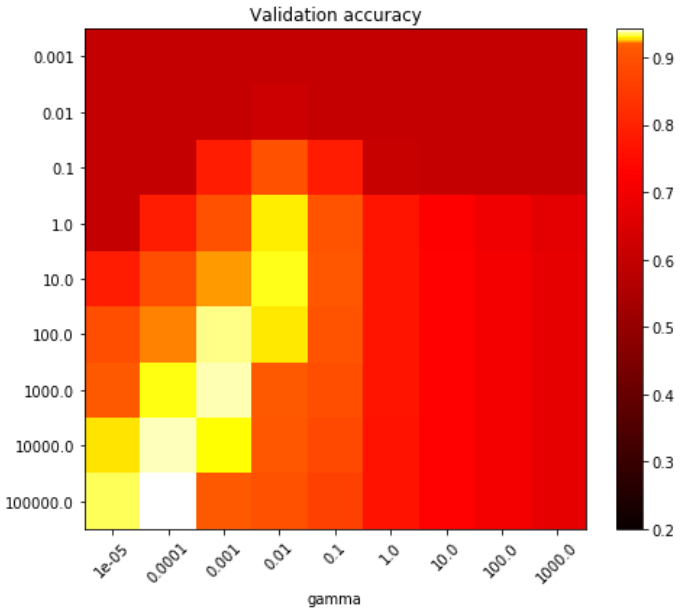
Output:

Clasifier	Train Accuracy	Validation Accuracy
Naive Bayes	0.8390	0.8363
Logistic Regression	0.9280	0.9214
linear SVC	0.9290	0.9202
SVC with linear kernel	0.9351	0.9267
Initial SVC with RBF kernel	0.9471	0.9286
Tuned RBF SVC	0.9642	0.9382
Decision Tree with max depth 10	0.9713	0.9134
Decision Tree with max depth 15	0.9852	0.9127
Adaboost Decision tree with depth 5	0.9997	0.9444
Adaboost Decision tree with depth 10	0.9997	0.9509
Adaboost Decision tree with depth 15	0.9997	0.9453
Random Forest	0.9892	0.9481

2.2 Grid Search for SVM

Generally for SVM, the RBF kernel is reasonable to be the first choice. Unlike the linear kernel, the RBF kernel can handle the case when relation between attributes is nonlinear. Grid searching of

parameters (*C* and *gamma*) with cross validation significantly contributes to satisfactory performance of SVM with RBF kernel. As shown in the heat map, in a coarse grid searching, the best SVM is obtained with parameters *C* = 100000.0 and *gamma* = 0.0001. It achieves the accuracy of 96.42% on training set and 93.82% on validation set. It improves a little bit accuracy compared to initial SVC with RBF kernel and parameters *C*=1 and *gamma* = 1/ # of features.




```
The best classifier is:  SVC(C=100000.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma=0.0001, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
```

Output:

In addition, a finer tuning can be achieved but at a much higher cost. And the quality of the resulting classifier may not have much improvement in this case.

2.3 Ensemble

Ensemble models have overall better performance than a single weak model. Through experiment, the quality of classification for decision tree methods varies as the depth of the tree changes. It threatens to overfitting problem if it grows too deep. Instead, Adaboost can alleviate the problem with weak classifiers but a strong result, 99% and 95% accuracy for training and validation respectively.

Another useful ensemble model is the random forest classifier. By tuning some of the parameters such as criterion and number of estimators, it can achieve pretty good result compared to previous classifiers.

n_estimators	criterion	min_samples_split	Train Accuracy	Validation Accuracy
30	entropy	10	0.9877	0.9478
30	gini	10	0.9864	0.9466
50	entropy	10	0.9888	0.9494
100	entropy	10	0.9892	0.9519
100	entropy	20	0.9803	0.9457
100	entropy	2	0.9997	0.9509

3. Conclusion

In summary, with 10-fold cross validation, all the candidate classifiers have better accuracy than the baseline classifier. Apart from ensemble methods, SVM with nonlinear kernel shows around 0.18 percent higher accuracy than SVM with linear kernel and logistic regression model. Ensemble models have better performance than single models overall. The best model that can be achieved in the experiment is Adaboost method with decision tree of max depth equaling to 5. Thus, the final step is to train this model with full training data and classify the testing data.