Report of Project 1

He Yajing 20459312

November 12, 2017

1 Introduction

For this project, I think I can use many kinds of classifiers to classify the training data. But they have different training accuracies. Thus, I consider it is significant to identify the classifier with a very high accuracy. Therefore, the process of my project includes data reading, classifier selection, model training and label prediction.

2 Data Reading

2.1 Training Data

For training process, there are two parts of data, one is training data and the other is training label. I read training data line by line and convert them into a numpy array with normalization. Also, for training label, I read it line by line and convert into a numpy array. Instead of normalization, I change the label from string type into int type.

```
#read training data
train_data = []
with open("traindata.csv") as rd:
    data = rd.readlines()
for line in data:
    train_data.append(line.split(","))
train_data = np.array(train_data)
train_data = preprocessing.normalize(train_data)
#print(train_data)

#read label data
label_data = []
with open("trainlabel.csv") as ld:
    label = ld.readlines()
label_data = np.array(label)
for i in range(len(label_data)):
    if label_data[i] == '0.0\n':
        label_data[i] = 0
    else:
        label_data[i] = 1
label_data = label_data.astype(np.int32)
#for i in range(len(label_data)):
 #   print(label_data[i])
```

Figure 1: Read training data and training label.

2.2 Test Data

For test data reading, it is similar to training data. I read the file line by line at first and convert into the numpy array with normalization.

3 Classifier Selection

After data is ready, I will firstly use training data to test different models and their own accuracies. Therefore, I choose the SVM model, logistic regression model, bagging classifier, gra-

Figure 2: Read test data.

dient boosting classifier, decision tree classifier, random forest classifier, MLP classifier, ensemble learning and so on. For some models, I also need to adjust different parameters to make the model highly accurate. Here comes the different accuracies for different models.

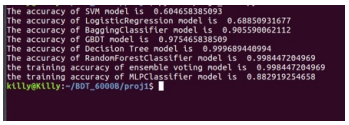

Figure 3: Accuracies of different classifiers

When I use ensemble model, it will take more than one classifiers to train the data and combine all their responses to generate the final results, which will improve the accuracy significantly. Therefore, I try to use ensemble model to combine the three classifiers with high accuracies. Among lots of possible combinations, I find that if I combine Logistic Regression, Random Forest and Decision Tree models and use ensemble voting model decide the final results, the accuracy will be extremely high. Also, it is important to adjust the weight of each classifier in ensemble model, so I try to put larger weight for more accurate classifier and higher accuracy turns out. Finally, I choose ensemble voting including Logistic Regression, Random Forest and Decision Tree models to do this project.

4 Model Training and Prediction

After I decide what model to use, I use normalized training dataset to train the model. For programming, there is a package called scikit-learn which include many classifiers and I can import needed classifiers as well as adjust parameters to get better performance. Finally, I get a model with 0.9997 accuracy. Then I use this model to predict test data by using predict function. This function will compute the probabilities of classification and put test data into a class with larger probability.Eventually, I put all predictions into a csv file to record them. The following is the code for classification model including weights for all classifiers.

Figure 4: Classification Model