Project 1 Report

Zhan Yunzhen ID: 20472766

In this project, I used Support Vector Machine (SVM) to do the classification.

Firstly, we need to load the data into Pycharm:

```
tmp = np.loadtxt("traindata.csv", dtype=np.str, delimiter=",")
data = tmp.astype(np.float)
tmp1 = np.loadtxt("trainlabel.csv", dtype=np.str, delimiter=",")
label = tmp1.astype(np.float)
tmp2 =  np.loadtxt("testdata.csv", dtype=np.str, delimiter=",")
tdata = tmp2.astype(np.float)
```

After observing the size of training dataset, we know that the dataset is a 3220x57 matrix. That means there are 3220 observations and each of them contains 57 features. Before building the classification model, we need to select some data from the training dataset as a verification dataset. Consequently, we select 25% from the training dataset to verify the model, the remains for model training.

```
trainingdata = data[0:2415,:]  # To train the model
traininglabel = label[0:2415]
testdata = data[2415:3220, :] # To verify
testlabel = label[2415:3220]
```

Then, we can begin to build the SVM model. In this case, I chose linear kernel function because there are only two types of the dataset.

```
clf = svm.SVC(kernel='linear', C=0.1).fit(trainingdata, traininglabel)
```

After model training, we used the verification dataset to test the model. As the output shown, the classification model reaches an accuracy of 92.17%.

```
scoresvm = clf.score(testdata, testlabel)
print(scoresvm)
```

```
/Users/zhanyunzhen/anaconda3/bin/python /Users/zhanyunzhen/PycharmProjects/6000b1/q2.py
0.921739130435
```

According to the result, we know that this model is useful. So we took the test dataset as an input to the model and predict the labels.

```
result = clf.predict(tdata)
```

The prediction is already saved in the CSV file.