Project 1

1. Pre-processing

1.1 Read data

Using pandas to read 'csv' file, read all these three files

1.2 Convert the frame to its Numpy-array representation

Before we use model, we need to convert all the files the format from DataFrame to a Numpy array.

Then using ravel to flat train label.

1.3 Normalization

Train data do normalization, minus mean, the divided by Standard Deviation, and test data do the same.

2. Model selection

2.1 baseline- Logistic Regression

Using logistic regression, split data 0.8 train and 0.2 test, fit the model and get the result, using mean square error do evaluate metrics, get the accuracy is 0.92857

```
0.9285714285714286
```

2.2 SVM-Kernel selection

Using SVM, do Cross-Validation, set K=5 to fit the model, select kernel.

Kernel=Linear, get the accuracy

```
0 Fold Training Accuracy :0.934394, Test Accuracy: 0.937888
1 Fold Training Accuracy :0.933618, Test Accuracy: 0.920807
2 Fold Training Accuracy :0.939053, Test Accuracy: 0.922360
3 Fold Training Accuracy :0.939829, Test Accuracy: 0.919255
4 Fold Training Accuracy :0.929348, Test Accuracy: 0.923913
```

Kernel=Polynomial

```
0 Fold Training Accuracy :0.790373, Test Accuracy: 0.746894
1 Fold Training Accuracy :0.788432, Test Accuracy: 0.787267
2 Fold Training Accuracy :0.786102, Test Accuracy: 0.793478
3 Fold Training Accuracy :0.794643, Test Accuracy: 0.731366
4 Fold Training Accuracy :0.788432, Test Accuracy: 0.779503
```

Kernel= Gaussian

```
0 Fold Training Accuracy :0.949922, Test Accuracy: 0.923913
1 Fold Training Accuracy :0.946817, Test Accuracy: 0.922360
2 Fold Training Accuracy :0.947205, Test Accuracy: 0.936335
3 Fold Training Accuracy :0.945652, Test Accuracy: 0.931677
4 Fold Training Accuracy :0.947593, Test Accuracy: 0.931677
```

So we choose Gaussian kernel get the final result.