Objective

This project is intended to classify data entries in a csv file. The given labeled dataset includes 3 parts, training data in a .csv file including 58 columns of float numbers, training label in a separate file with 0 or 1 labeled, and a testing file waiting for corresponding labels.
This project is a classic classification problem, so here two classification models are used, SVM and Logistic Regression.

Process:

1. Load csv data into pandas data frame for processing.
2. Get the mean and standard deviation of the training and testing data, as a result:

```
In [116]: print 'The average of the training data is: ', np.mean(trainingData.mean(axis=0))
          print "The standard deviation of the training data is:",np.mean(trainingData.std(axis=0))

          The average of the training data is:  6.03357911627
          The standard deviation of the training data is: 12.5719911587
```

3. Normalize the data so as the standard deviation is 1.

```
In [119]: trainingData = preprocessing.scale(trainingData)
          testingData = preprocessing.scale(testingData)
          trainingData.mean(axis=0)
          trainingData.std(axis=0)

          print 'The average of the training data is: ', np.mean(trainingData.mean(axis=0))
          print "The standard deviation of the training data is:",np.mean(trainingData.std(axis=0))

          The average of the training data is:  5.13918209503e-18
          The standard deviation of the training data is: 1.0
```

4. Build 2 classification models using SVM and logistic regression, with k-fold cross-validation, where k=10 in this project.
5. Run the two models with training data to compare the result

```
In [128]: svmClassify(trainingData,trainingLabel,testingData)

          0.928881987578
```

```
In [129]: classifylr(trainingData,trainingLabel,testingData)

          0.921739130435
```

6. Post-processing: select the model with the best accuracy, which is 0.92888 for SVM and save the result to csv format.