MSBD 6000B Project I Report

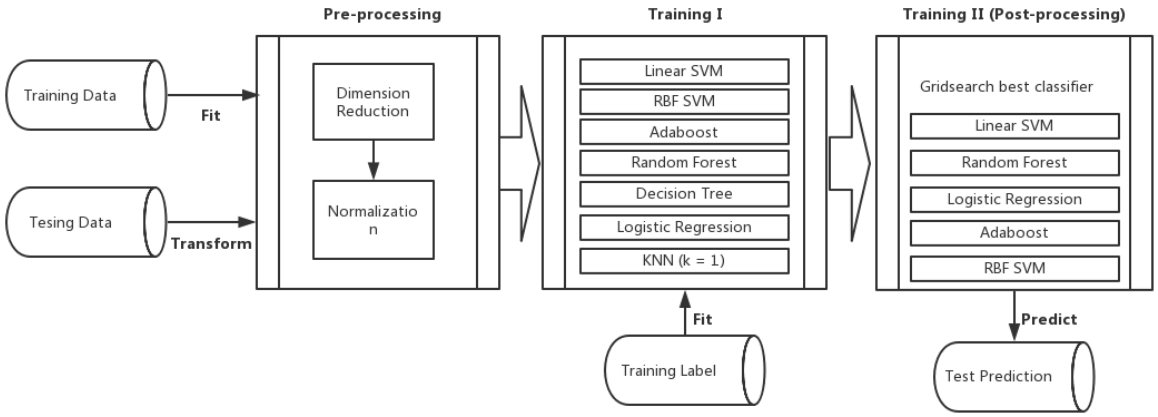
Student Name: LIU Yan Yun Student Id: 20384933

I. Introduction

In this project, I use a group of traditional machine learning classifiers which are well-packaged in sklearn to solve the given classification problem. The main workflow of my project is listed below：

1. Load the training data and training label
2. Dimension reduction.
3. Data normalization and standardisation.
4. Train machine learning models with seven classifiers.
5. Given seven output predictions as input features, select the best classifier from grisearch to give last-layer prediction.
6. Re-train the model with all the data and give prediction for test data.

The workflow is shown below:



II. Pre-processing and dimension reduction

Before we start - check correlation among features

To have a better understanding about the dataset, I check the correlation coefficient for features with `np.corrcoef`. The result indicates that a few features are highly linear dependent.

Dimension Reduction

To aviod collinearity, I use several feature selection tools to check which features need to be preserved. In function `Select_best_features`, I set five dimention reduction methods and use grid search to select one with the best score.

```
VarianceThreshold(threshold=(.8 * (1 - .8))),
PCA(32),
PCA(48),
SelectFromModel(LinearSVC(C=1, penalty="l1", dual=False)),
SelectFromModel(ExtraTreesClassifier())
```

Data standardization and normalization

Pre-processing stage is ensential in machine learning methodology. A standard, normalized dataset will provide some benefits for training. In function `Normalize_train_test_features`, I use `StandardScaler` in sklearn to fit-transform training data and transform testing data.

III. Training model

In traning stage, my idea is to train seven classifiers and output seven set of predictions. Then use the seven output as input features, again train another model to get final prediction.

Stage I: Using seven classifier to get predictions

To get more robust prediction, I use seven different classifiers and train these models one by one, predicting seven results for each test sample as new input features, and give prediction for each training sample as new training features as well. The seven classifiers are listed below:

```
model_svm = SVC(kernel= 'rbf', C=1, degree=2)
model_ada = AdaBoostClassifier(base_estimator=DecisionTreeClassifier())
model_lr = LogisticRegression()
model_lsvm = SVC(kernel= 'linear')
model_rf = RandomForestClassifier()
model_knn = KNeighborsClassifier(n_neighbors=1)
model_dt = DecisionTreeClassifier()
```

Stage II: Give final prediction using seven features

In this part, I used the output predictions as 7 features to predict final result. The model is selected by grid-search in function `Get_final_prediction_gridsearch`.

VI. Model Evaluation

For final result, the cross validation accuracy in training data is 95.0%, and the confusion matrix is shown below:

```
             precision    recall  f1-score   support

    class 0       0.95      0.97      0.96      1947
    class 1       0.95      0.93      0.94      1273

avg / total       0.95      0.95      0.95      3220
```