MSBD6000B Project 1 Chan Kin Ho 20411918

Preprocessing:

Each feature is standardized by subtracting the mean and divided by the standard deviation.

Model Selection:

Random Forests algorithm is chosen as the model. It is an ensemble method specifically designed for decision tree classifiers. Many trees are grown to classify a resample of the original data and their results are aggregated to give a final classification of the data. In particular, the algorithm consists of the bagging method and random vector method. Bagging means building tree on each bootstrap sample, and the forest chooses the classification result having the most votes by all trees. While random vector means at each node, the best split is chosen from a random subset of the features instead of using all.

To choose the hyperparameters of the random forest algorithm such as number of ensembles and the maximum number of features used for best split, a 10-fold stratified cross validation is implemented. Data is split into 10 folds so that the mean response value is approximately equal in each fold, which make the classifier learn both class in balance. One-fold is then taken out as validation set in turn and the rest is used to train the model. Hyperparameters are chosen based on the mean validation error of the 10 iterations.

The following hyperparameters are chosen:
Number of estimator: 200
Maximum number of features to consider when looking for the best split: 3
Minimum number of samples required to split an internal node: 3
Minimum number of samples required to be at a leaf node: 1

Mean Training accuracy: 0.9993
Mean Validation accuracy: 0.9550

Postprocessing and Prediction:

Each feature in testing data is standardized by subtracting the mean of the training data and divided by the standard deviation of the training data before being predicted by the model. The model is fitted using all the 10 folds of training data instead of the 9 folds in model selection stage.