EMIS 3309: Information Engineering

Lab 7

For this lab, you will need to go to 'UCI Machine Learning Repository' website. Use the link provided below and download 'winequality-red.csv'. Read description on the website to understand features and response values stored in the dataset.

Source of the dataset: `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`

1. Import data from the downloaded file to R. Save the dataset as wine.

2. Set seed of 100 for the random number generator, then split 'wine' into training set (containing 1000 samples) and testing set (containing the remaining samples).

3. Apply KNN **regression** method using 3-nearest neighbors to make predictions on the testing data; use all feature values to predict 'quality' of wine. What is the mean squared error (MSE)?

4. Since there are many options for choosing the number of nearest neighbors, we are going to perform cross validation on the **training data** with:

   - Number of neighbors: 1, ..., 30;
   - Number of folds: 5 (i.e., we are going to use 5-fold cross validation method).

   At the end of all runs, report the **average** MSE for each value of the number-of-neighbors.

5. Create a line plot for the above result. Let x-axis indicate the number of neighbors and y-axis indicate the average MSE.

6. Based on the result of cross validation, choose a value for the number-of-neighbors. Apply KNN method with the number-of-neighbors you picked. Report MSE of the method on the testing data.

7. Using training data, fit a linear model to predict 'quality' of wine. Compute MSE of the model on the testing data.

8. Each of KNN and linear model makes certain assumptions on the distribution of data points. Based on the result of previous problems, make an inference on why one method works better than the other method for the wine dataset.