Machine Learning : Scikit - Learn

Question 1 (Classification)

(i) Read the classification datasets (training and test sets) into a NumPy array. This is a binary classification problem. As usual the last column is the class label (0 or 1). All other columns are numerical features.

(ii) Using Scikit Learn build a basic kNN classifier model for this dataset (start with k=1) and assess its classification accuracy.

(iii) Explore the impact of adopting various values of k on your model and different distance metrics.

(iv) Next contrast the performance of the kNN model with other classification models. For example, try the following:

a. DecisionTreeClassifier

b. Naïve Bayes

c. SVM

Question 2 (Regression)

(i) In the exercise folder you will find a file called regressionExample.csv. Read this file into a NumPy array

(ii) Use train_test_split to split the dataset into 20% test and 80% training.

(iii) Use a KNeighborsRegressor to build a predictive model for the data and assess its accuracy using R2.

Question 3 (Outlier Detection)

(i) In the exercise folder you will find a zip file called outlierData.zip. This zip file contains a training file called trainOutlier.csv and a test file called test.csv. This is a regression problem and target value is contained in the last column in each file. Read this data into your program.

(ii) Build a model using DecisionTreeRegressor and assess the accuracy (using R2).

(iii) Identify any outliers in the training data using boxplots.

(iv) Remove the outliers and reassess the new accuracy of the model.

(v) Try an alternative model such as a KNeighborsRegressor and record the accuracy.