

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 (Outlier Detection for Regression)

- (i) In the exercise folder you will find a zip file called outlierData.zip. This zip file contains a training file and test csv file. 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 [Lasso Regression](#) and assess the accuracy (using R2).
- (iii) Use an [Isolation Forest](#) to identify multi-variate outliers in feature space.
- (iv) Remove the identified outliers and reassess the new accuracy of the model.
- (v) Try alternative model such as a [KNeighborsRegressor](#) or [DecisionTreeRegressor](#). Notice the sensitivity of different models to the outliers.