Summary Report

Assignment 1

Advanced Machine Learning_64061

MSBA

Submitted by: Dileep Kumar Pasala

Introduction:

In this work a synthetic dataset is generated by utilizing the K-Nearest Neighbours (KNN) classification approach and subsequently validated the performance by exploring the results and assessed the classifier's ability in recognizing and differentiating the data points by feature similarity.

Analysis:

Synthetic dataset with 150 samples is generated through scikit-learn's make_blobs method which is used to create data for clustering & classification purposes. The data set is segregated into three separate classes (Class 0,1,2 respectively) to represent a specific pointed cluster in 2-D space as follows [2,4], [6,6], [1,9]]. Subsequently the dataset is split into training and testing sets on 80:20 ratio respectively. A visualization of graph plots for training data set and testing data set was created in which the train plot shows 3 unique clusters for each class and test plot shows same pattern as train data plot. For model training KNN classification method by Euclidean distance metric is utilized and predictions were generated to study model performance by the metrics like accuracy score and confusion matrix. An accuracy of 100% is seen for the training and testing. Similarly, confusion matrix categorized the test samples without misclassification and the classes 0,1,2 was predicted accurately (for class 0: 14, class 1:8, class2:8 data points were correctly classified respectively) which shows differentiating capability of the model between clustered created.

Conclusion: 100% accuracy for the train and test data sets predictions through KNN classification.