DS-326 Machine Learning for Data Science

Assignment 1

Name: …….. Due Date:30-1-2023

Roll#: …….. Total Marks: 50

Dataset Description:

The time domain dataset includes the instantaneous values of three-phase current signature data of loaded induction motor with healthy and unhealthy conditions. The motor with various faults were operated under different load conditions to observe their impact on current signatures. The motor faults include inner-race and outer-race bearing faults of (i) 0.7mm, (ii) 0.9mm, (iii) 1.1mm, (iv) 1.3mm, (v) 1.5m, and (vi) 1.7mm. The bearings with different severity levels of faults were operated under the load conditions of 100W, 200W, and 300W. In addition, the motor health condition was also experimented under broken rotor bar (BRB) fault with 100W and 300W loads. There are a total 39 datasets including the three-phase current data acquired under different motor health conditions. The data was acquired at the sampling rate of 10 kHz at the rate of 1000 samples per channel using non-invasive current sensors. Seven folders containing the different induction motor datasets on different loads.

Note: This is 14 class classification problem i.e., healthy, unhealthy, unhealthy due to bearing faults (inner and outer race) and unhealthy due to rotor broken fault on different loads.

Question#1: You have to implement K-NN algorithm on the above-mentioned dataset with the following steps:

- Data preprocessing: In the dataset, the total number of samples in each file are more than 100,000 and any block of 1000 samples can be used to label the corresponding healthy and unhealthy output. Create a new csv file for training and testing your model in which each row containing 1000 samples with assignment of the output label healthy/unhealthy.
- Calculate the Euclidean distance
- k-NN Model Implementation without build-in libraries.
- Apply hold-out and 10 cross validation method for training the model.
- Calculate the classification measures. (Accuracy, Recall, Precision, Sensitivity, specificity, F1 Score)
- Find the optimized K value and plot cross-validation and test accuracy based on the K values. You can tune your model on multiple K iterations.

Question#2: Apply the data dimensionality reduction technique PCA on the dataset and implement all the subparts of question1.

Question#3: Which model gives the best accuracy and why?