WT Lab Evaluation Test

Problem: Write a program to implement Minimum Distance Classifier algorithm to classify the iris dataset. Print all the intermediate results of every step and finally accuracy of the classifier.

Minimum Distance Classifier (MDC)

1. MDC is simple classification algorithm used for classification problems.
2. MDC is basically classified Test Samples based on similarities (minimum distance) with Training Samples.
3. Calculate the Euclidean distance with all the training samples (points) to one given test sample (point) and select the lowest distance sample (point).
4. Then, identify or predict the Class Label of the test sample is decided based on the Class Label of the minimum distance training sample.
5. Similar way follows the step 3 and 4 to identify or predict the Class Labels of all test samples (points).
6. Finally, compare the actual Class Labels of the test samples with the evaluated or predicted Class Labels of all test samples in the experiment.
7. Compute the correctly-classified test samples based on the given condition as **the estimated (evaluated) Class Levels are same as desired (actual) Class Labels of the test samples**. Then, find out Accuracy of the Classifier in terms of percentage as follows.

$$\text{Accuracy (\%)} = \frac{Correctly\ Classified\ Test\ Samples}{Total\ Test\ Samples} \times 100$$

Implementation Steps:

(i) Read Iris dataset "iris.csv" file which contains 150 iris flower Samples and each sample has 4 features. In addition, each Sample is associated with a Class Label.

(ii) Normalize the dataset (feature values) within the range of 0 to 1 using Min-Max Normalization.

(iii) Make three groups (Classes) namely "Setosa", "Virginica", and "Versicolor" contain 50 Samples each. In other words, create a 2D array for each class of size (50 x 4) that contains 50 Samples and each Sample has 4 features.

(iv) Divide entire dataset into two sets namely Training set and Test set on the basis of the percentage of training samples input from keyboard. As an example, from 150 dataset samples, Training set contains 70% i.e., 105 Samples and remaining 30% i.e., 45 samples are in test set. In addition, 70% Training Samples means 70% from every classes i.e., 70% out of 50 Samples is 35 samples from each class in total 105 samples. Similar, 30% Test samples means 30% from every classes i.e., 30% out of 50 samples is 15 Samples from each class in total 45 Samples.

(v) Once Training and Test sets are obtained, then compute Euclidean distance between each test sample and all Training Samples. That means from the above example, for each Test Sample find out Euclidean distance between 105 Training Samples respectively. Finally, compute Distance matrix i.e., 2D

array of size (45 x 105), which contains 45 rows of Test samples and their Euclidean distance values with respected to 105 columns of Training Samples.

(vi) Find out minimum Euclidean distance for each Test Sample which is a Training Sample in the Distance matrix. That means, a particular Test Sample (index *i*) is as much as similar to a Training Sample (index *j*).

(vii) Check each Test Sample (index *i*) Class Label with Training Sample (index *j*) Class Label, if they matched then Test Sample is correctly-classified otherwise miss-classified. Finally, count a number of Test Samples correctly-classified out of total Test Samples. For example, let say 40 Test Samples are correctly-classified from 45 Samples, then compute the performance of the classifier based on Accuracy: $\frac{40}{45} \times 100 = 88.88\%$.