CP 3252- Machine learninng Lab

1. Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 2 clusters: A1=(2,4), A2=(2,9), A3=(8,3), A4=(5,7), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Write the distance matrix based on the Euclidean distance.

Suppose that the initial seeds (centers of each cluster) are A2, A3 and A7. Run the k-means algorithm for 2 epochs only. At the end of this epoch show:

a) The new clusters (i.e. the examples belonging to each cluster)

b) The centers of the new clusters

c) Draw a 10 by 10 space with all the 8 points and show the clusters after the second epoch and the new centroids.

d) Implement and show how many more iterations are needed to converge? Draw the result for each epoch.(only implementation for Qn.d)

2. For the given dataset,

Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

1.Implement KNN Classifier on the given dataset to predict the species for the given flower.

Sepal Length	Sepal Width	Species
5.2	3.1	?

Calculate the distance between the query instance and all the training samples using Euclidean distance to find the k nearest neighbour for the given flower.

1. Use the simple majority of the category of nearest neighbors as the prediction value of the query instance. (Use k-3,k=4 and k=5).
2. Analyse whether the predicted species is same if the distance metric used is Manhattan distance(|x1-x2|+|y1-y2|)

3. Construct Decision Tree and find accuracy of the classifier for the given dataset.

TASK

The customer Dataset (loan.csv) consists of information about 381 customers and status of their loan application result (Y/N) as a binary (2-class) classification problem.

a) Consider the columns: Gender, Married, Education, ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term, Credit_History, Property_Area as feature Input (X) and Loan_Status as label ouput (Y).
b) Perform the necessary conversion (label encoding and feature scaling) of appropriate features.
c) Classify the data set (20% test, 80% training) by using both Decision Tree.
d) For the Decision Tree Algorithm generate the Tree and show/save it as an image file.
f) Show the accuracies and confusion matrices for the test set.