PESIT Department of Computer Science and Engineering

Course: Data Mining
Semester: 2016 Spring (January – May)
Instructor: BNR (Dr. B. Narsing Rao)

Assignment: 01
Topic: Object Similarity and Classification
Due by: Midnight on **Tuesday, January 12, 2014**
Method: Send zip archive (.zip, .rar, etc.) by email to bnrao@pes.edu
The name of the zip archive should be: DM-A01-your USN-your name
(USN must be upper case and your name should be in mixed case)
The zip archive should contain the following (see below for details):

1. Program (in a text file)
2. Output from program (in a text file)

This assignment has to be done using the file **iris.csv** (provided) which contains data on 150 iris plants. Each plant has the following attributes: ID, Sepal Length (cm), Sepal Width (cm), Petal length (cm), Petal Width (cm), class (Setosa, Versicolour, Virginica).

Write a program that read the data from the iris.csv file and carry out the following steps:

Step 1

Partition the data randomly into two sets as follows:
Training Set (140), Test Set (10)

Print out the IDs of the objects in the test set.

Step 2 (use Training Set only)

1. Compute and display summary statistics (Minimum, Maximum, Mean, Median, and Standard Deviation) for each numerical attribute for the whole data set as well as by class

2. Generate the following output using the appropriate distance computations (Note: use only the four numeric attributes for this purpose):

Distance measure	**Euclidean**	**Manhattan**	**Supremum**
Minimum Value			
Maximum Value			

For each value in the above table, list also the ID and class of the objects it corresponds to. For example, suppose the minimum between any two instances was 0.5, and it corresponded to the distance between 23 and 109, the listing should say something like:

0.5, 23 (setosa), 109 (virginica)

Step 3 (Use both Training and Test Sets)

Predicts the class of each of the Iris plants in the test set using the k-Nearest-Neighbor approach:

1. Accept a test set of attributes from the test set
2. Determine the k nearest neighbors using each of the three distance measures
3. Predict the class based on the class of the neighbors
4. Repeat the steps above for values of k between 1 and 5 both inclusive

The output from the program should be as follows:

k =

		Predicted Class using		
Test ID	Actual Class	Manhattan	Euclidean	Supremum
…				

Please note the following:

- There should be one table as shown above for each value of K
- Each table should have 10 rows corresponding to each of the 10 test objects
- Each entry in the table for the predicted class should contain the classes of the k nearest neighbors and the class finally chosen. For example, if k=3, the entry might look something like this:
    - Nearest neighbors: setosa, virginica, setosa
    - Predicted class: setosa