Laboratory Manual

Subject: DWM Lab | Class/Sem / Batch: T.E. /V

Name of Faculty: Prof. Archana Kotangale | Department: Computer Engineering

Academic Year: 2022-2023

Experiment – 8

Title: Implementation of Clustering algorithm (K-means).

Objective:

• To learn how to classify data by K nearest neighbor algorithm for classification

Reference:

• Data Mining Introductory & Advanced Topic by Margaret H. Dunham

• Data Mining Concept and Technique By Han & Kamber

Pre-requisite:

• Fundamental Knowledge of Database Management

Theory:

Ink-nearest-neighbor classification, the training data set is used to classify each member of a "target" dataset. The structure of the data is that there is a classification (categorical) variable of interest ("buyer, "or" non-buyer, "for example), and a number of additional predict or variables (age, income, location...).

Algorithm:

1. For each row (case) in the target data set (the set to be classified), locate the k closest members (the k nearest neighbors) of the training data set. A Euclidean Distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
2. Examine the k nearest neighbors-which classification (category) do most of them belong to? Assign this catego y to the row being examined.
3. Repeat this procedure for the remaining rows (cases) in the target set.
4. Also lets the user select a maximum value for k, builds models parallel on all values of k up to the maximum specified value and scoring is done on the best of these models. The computing time goes up as k goes up, but the advantage is that higher values of k provide smoothing that reduces vulnerability to noise in the training data. In practical applications, typically, k is in units or tens rather than in hundreds or thousands.

Laboratory Manual

Subject: DWM Lab | Class/Sem / Batch: T.E. /V

Name of Faculty: Prof. Archana Kotangale | Department: Computer Engineering

Academic Year: 2022-2023

Input:

Name	Gender	Height(m)
Kristina	F	1.6
Jim	M	2
Maggie	F	1.9
Bob	M	1.85
Dave	F	1.7
Kimm	M	1.9
Todd	M	1.9
Amy	F	1.85
Kathy	F	1.6

2m<=Tall, 1.7m< H<2m Medium, H<=1.7m Short

New Tuple <Pat,F,1.6> , suppose K=5 is given than K nearest neighbors to input tuple {(Kristina,F,1.6) ,

(Kathy,F,1.6),(Dave,F,1.7)}

Output: Pat – Short

Conclusion:

K- means clustering is simplest method used for forming data clusters