**MACHINE LEARNING – Project 3**

**K-means Clustering Report**

**CLUSTERING:**

Clustering is a Machine Learning technique for grouping of set of unlabeled data points into a specific group/ cluster. It is basically a type of unsupervised learning method. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples. The data points are in one cluster should have similar properties and features while data points of different cluster must have dissimilar features. It is a collection of objects on the basis of similarity and dissimilarity between them.

**K-means ALGORITHM:**

The clustering problem is solved by an algorithm called **K-means Algorithm** which is a unsupervised, non-deterministic and iterative algorithm. K-means clustering is one of the simplest and popular unsupervised machine-learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes. The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

A cluster refers to a collection of data points aggregated together because of certain similarities. You'll characterize an objective number k, which alludes to the quantity of centroids you need in the dataset. A centroid is the nonexistent or genuine area addressing the focal point of the bunch.

Each information point is designated to every one of the bunches through lessening the in-group amount of squares. At last, the K-means algorithm identifies k number of centroids, and afterward designates each data point to the closest cluster, while keeping the centroids as little as could really be expected. The ‘means’ in the K-means refers to averaging of the data, i.e., finding the **centroid**.

**WORKING OF K-means ALGORITHM:**

To process the learning data, K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative calculations to optimize the positions of the centroids. It stops creating and optimizing clusters when either:

* The centroids have stabilized – there is no change in their values because the clustering has been successful.
* The defined number of iterations has been achieved.

**Steps of K-means Algorithm:**

Step 1 - Choosing the number of clusters “k”

Step 2 – Randomly choose the “k” centroids and assign each point to a cluster with closest centroid.

Step 3 – Repeat the following until centroids stop changing:

* For each cluster, compute the cluster centroid by taking the mean vector of the points in the cluster.
* Assign each data point to the cluster for which the centroid is the closest.

**Finding the optimal value of K:**

The Elbow method is used for optimizing k value.

**Inertia**: It is a boundary which computes the sum of distances of the multitude of points inside the cluster from the centroid of the cluster. To have the similar points in a same cluster the distance between the points should be low. Lesser the value of inertia, better the clusters.

Chart, line chart

Description automatically generated

When a graph is plotted between the inertia and K values, **the value of K at which elbow forms gives the optimum**.

**Implementation of K-means from Scratch:**

1. **Import Libraries**

**Graphical user interface, text, application

Description automatically generated**

1. **Read .CSV file**

**Graphical user interface, text, application

Description automatically generated**

Here we are splitting the data into train data as list of list and then in label it is splitting the data into list of labels containing the values 0,1,2A picture containing table

Description automatically generated

1. **Initializing number of clusters “k” and iterations**

**Text

Description automatically generated**

1. **Defining the distance**

**Text

Description automatically generated**

1. **Defining K-means Function**

**A picture containing text

Description automatically generated**

1. **Applying K-means and visualize the output**

**K-means results for Euclidean Distance**

**Text

Description automatically generated**

**K-means results for Manhattan Distance**

**Text

Description automatically generated with medium confidence**

**K-means results for Minkowski Distance**

**Text

Description automatically generated with medium confidence**

**K-means results for Cosine Distance**

**A picture containing text

Description automatically generated**

**Accuracy with different features with respect to Euclidean Distance**

**Text

Description automatically generated**

**Elbow curve with the Euclidean Distance and Manhattan Distance with different number of clusters**

**Chart, line chart

Description automatically generated  
Chart, line chart

Description automatically generated**

**Accuracy curve with the Euclidean Distance and Manhattan Distance with different number of clusters**

**Chart, line chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**CONCLUSION:**

From the above experiment it is clear and evident that when we keep the cluster number to three we are getting the maximum accuracy and the distance which we get the maximum accuracy is the Manhattan with an accuracy of 89% and the next distance which is best is the Euclidean with an accuracy of 88% and we need to consider all the features in order to get a best accuracy when we consider 2 features the accuracy is less

**REFERENCES:**

[**https://pythonprogramming.net/k-means-from-scratch-2-machine-learning-tutorial/**](https://pythonprogramming.net/k-means-from-scratch-2-machine-learning-tutorial/)

[**https://stackoverflow.com/questions/54845820/k-means-in-python-from-scratch/54845896**](https://stackoverflow.com/questions/54845820/k-means-in-python-from-scratch/54845896)