**ASSIGNMENT HELP**

**MANUAL**

**ASSIGNMENT-4**



SUBMITTED

TO

VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY, PUNE

FOR THE SKILL AND COMPETENCY EVALUATION OF

**DATA SCIENCE & MACHINE LEARNING**

IN

**CSE AI DEPARTMENT**

BY

**Divyansh Mohta [22210998]**

**Class: S.Y. BTech Division: A Batch: A2**

**Batch Teacher**

**Dr. ANURADHA YENKIKAR.**

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Contents** | **Page No.** |
| **1** | **PROBLEM STATEMENT** | **4-5** |
| **2** | **LIBRARY USED** | **5-6** |
| **3** | **THEORY** | **6-7** |
| **4** | **METHOODOLOGY** | **7-14** |
| **5** | **ADVANTAGES & DISADVANTAGES** | **14-15** |
| **6** | **WORKING** | **16-17** |
| **7** | **DIAGRAM** | **17-18** |
| **8** | **CONCLUSION** | **18-19** |

1.PROBLEM STATEMENT:

Write a program to do following:

We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16,

0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering

with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2.

Answer the following:

a) Which cluster does P6 belong to?

b) What is the population of a cluster around m2?

c) What is the updated value of m1 and m2?

2. LIBRARY USED:

* scikit-learn (sklearn): Utilized for machine learning tasks such as splitting the dataset into training and testing sets, scaling features, building a decision tree classifier, and evaluating the model's performance using classification metrics.

3. THEORY:

K-Means Clustering is an Unsupervised Machine Learning algorithm, which groups the unlabeled dataset into different clusters. The article aims to explore the fundamentals and working of k mean clustering along with the implementation.

K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the cluster based on its distance from centroid of the cluster. After assigning each point to one of the cluster, new cluster centroids are assigned. This process runs iteratively until it finds good cluster. In the analysis we assume that number of cluster is given in advanced and we have to put points in one of the group.

The algorithm works as follows:

we randomly initialize k points, called means or cluster centroids.

We categorize each item to its closest mean, and we update the mean’s coordinates, which are the averages of the items categorized in that cluster so far.

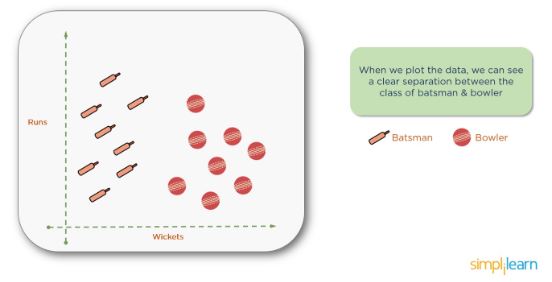
We repeat the process for a given number of iterations and at the end, we have our clusters.

4. METHODS:

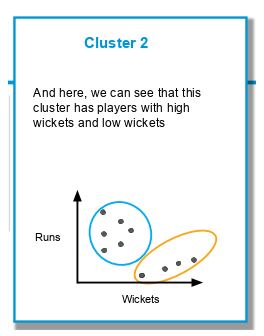
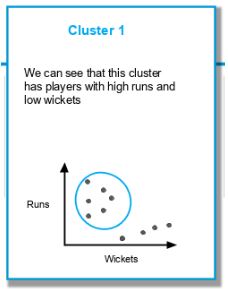
Assign data points

Here, we have our data set plotted on ‘x’ and ‘y’ coordinates. The information on the y-axis is about the runs scored, and on the x-axis about the wickets taken by the players.

If we plot the data, this is how it would look:



We need to create the clusters, as shown below:



The first step in k-means clustering is the allocation of two centroids randomly (as K=2). Two points are assigned as centroid

The next step is to determine the distance between each of the randomly assigned centroids' data points. For every point, the distance is measured from both the centroids, and whichever distance is less, that point is assigned to that centroid.

The next step is to determine the actual centroid for these two clusters. The original randomly allocated centroid is to be repositioned to the actual centroid of the clusters.

This process of calculating the distance and repositioning the centroid continues until we obtain our final cluster. Then the centroid repositioning stops.

5. ADVANTAGES AND DISADVANTAGES:

Here are the advantages and disadvantages of the k-means clustering system:

Advantages:

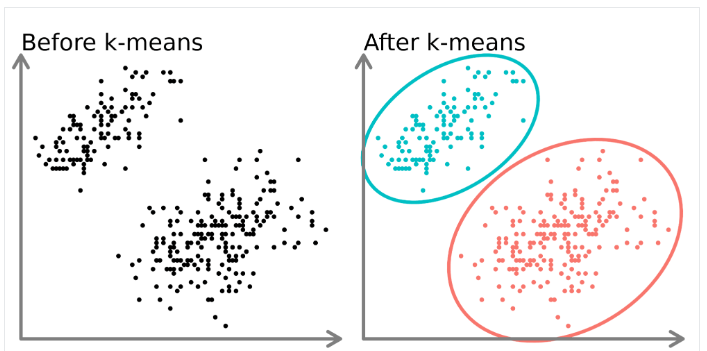
1. Simple and easy to implement: The k-means algorithm is easy to understand and implement, making it a popular choice for clustering tasks.
2. Fast and efficient: K-means is computationally efficient and can handle large datasets with high dimensionality.
3. Scalability: K-means can handle large datasets with a large number of data points and can be easily scaled to handle even larger datasets.
4. Flexibility: K-means can be easily adapted to different applications and can be used with different distance metrics and initialization methods.

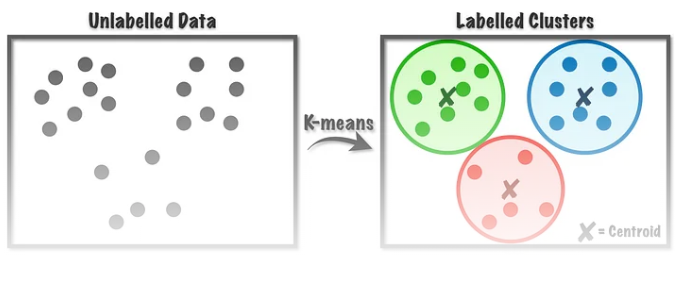
Disadvantages:

1. Sensitivity to initial centroids: K-means is sensitive to the initial selection of centroids and can converge to a suboptimal solution.
2. Requires specifying the number of clusters: The number of clusters k needs to be specified before running the algorithm, which can be challenging in some applications.
3. Sensitive to outliers: K-means is sensitive to outliers, which can have a significant impact on the resulting clusters.

Overall, while k-means clustering is a powerful and widely-used clustering algorithm, it is important to consider its limitations and potential drawbacks when applying it to real-world datasets.

**6. DIAGRAM: -**





SOME VISUALIZATION

**7. CONCLUSION:**

In conclusion, K-means clustering is a powerful unsupervised machine learning algorithm for grouping unlabeled datasets. Its objective is to divide data into clusters, making similar data points part of the same group. The algorithm initializes cluster centroids and iteratively assigns data points to the nearest centroid, updating centroids based on the mean of points in each cluster.