**Assignment Clustering**

**V SURESH KUMAR**

**Assignment No : 04**

1) What is Clustering/ Segmentation?

Segmenting is the process of putting customers into groups based on similarities, and clustering is the process of finding similarities in customers so that they can be grouped, and therefore segmented. They seem quite similar, but they are not quite the same.

Example : Segmentation

When you segment you know who to target. If I’m selling an expensive little black dress, I want to target women who have a high annual household income. In this case, I’m defining the limits of the group. Women. With annual incomes over a hundred thousand dollars who have purchased similar items in that product category. It’s natural to assume that this group of women would (be able to) buy my store’s dresses.

Example : Clustering

Clustering is the process of using machine learning and algorithms to identify how different types of data are related and creating new segments based on those relationships.

Clustering finds the relationship between data points so they can be segmented

2) How do you determine the "nearness" of clusters?

Types of clusters

• Well-separated clusters

• Center-based clusters

• Contiguous clusters

• Density-based clusters

• Conceptual clusters

Well-Separated Clusters:

A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.

Center-based clusters :  
 A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster .  
 The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster.

Contiguity-based clusters :  
 Contiguous Cluster (Nearest neighbor or Transitive) – A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster

Density-based :  
 Density-based – A cluster is a dense region of points, which is separated by low-  
density regions, from other regions of high density. Used when the clusters are irregular or intertwined, and when noise and outliers are present.

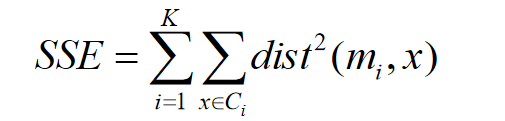
Conceptual clusters :  
 Shared Property or Conceptual Clusters – Finds clusters that share some common property or represent a particular concept.

Algorithm to cluster types mapping   
 K-means and its variants   
1.Center-based   
2.Density-based   
 DBSCAN clustering   
1.Density-based   
2.Contiguity-based

**K means clustering** :  
• Exclusive clustering approach   
• Each cluster is associated with a centroid (center point)   
• Each point is assigned to the cluster with the closest centroid   
• Number of clusters, K, must be specified   
• The basic algorithm is very simple

Details   
• Initial centroids are often chosen randomly.   
– Clusters produced vary from one run to another.   
• The centroid is (typically) the mean of the points in the cluster.   
• ‘Closeness’ is measured by Euclidean distance, cosine similarity,   
correlation, etc.   
• K-means will converge for common similarity measures   
mentioned above.   
• Most of the convergence happens in the first few iterations.   
– Often the stopping condition is changed to ‘Until relatively few points   
change clusters’   
• Complexity is O( n \* K \* I \* d )   
– n = number of points, K = number of clusters,   
I = number of iterations, d = number of attributes.

How to measure bad?

• Most common measure is Sum of Squared Error (SSE)   
– For each point, the error is the distance to the nearest cluster   
– To get SSE, we square these errors and sum them.   
  
– x is a data point in cluster Ci and mi is the representative point for cluster Ci   
• can show that mi corresponds to the center (mean) of the cluster   
– Given two clusters, we can choose the one with the smallest error   
– One easy way to reduce SSE is to increase K, the number of clusters   
• A good clustering with smaller K can have a lower SSE than a poor clustering with higher K.

How to fix?   
• Multiple runs   
–Helps, but probability is not on your side   
• Select more than k initial centroids and then select among these initial centroids   
• Select most widely separated initial centroids   
• Bisecting K-means   
–Not as susceptible to initialization issues.

Bisecting K means   
• Bisecting K-means algorithm – Variant of K-means that can produce a partitional or a hierarchical clustering.

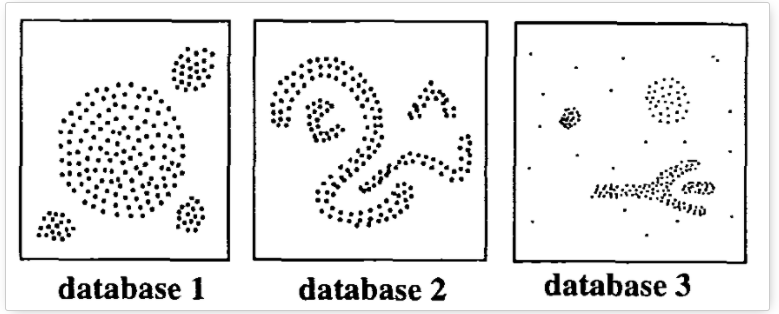
**DBSCAN :**

• DBSCAN is a density-based algorithm. – Density = number of points within a specified radius (Eps)   
– A point is a core point if it has more than a specified number of points (MinPts) within Eps   
• These are points that are at the interior of a cluster   
– A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point   
– A noise point is any point that is not a core point or a border point.

# Concepts of density-based clustering:

**Partitioning methods** (**K-means**, **PAM** **clustering**) and **hierarchical clustering** are suitable for finding spherical-shaped **clusters** or convex clusters. In other words, they work well for compact and well separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data.

The basic idea behind **density-based clustering** approach is derived from a human intuitive clustering method. For instance, by looking at the figure below, one can easily identify four clusters along with several points of noise, because of the differences in the density of points.



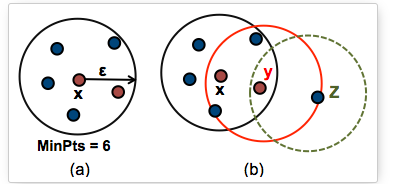
# Algorithm of DBSCAN:

The goal is to identify dense regions, which can be measured by the number of objects close to a given point.

Two important parameters are required for **DBSCAN**: **epsilon** (“eps”) and **minimum points** (“MinPts”). The parameter **eps** defines the radius of neighborhood around a point x. It’s called called the \(\epsilon\)-neighborhood of x. The parameter **MinPts** is the minimum number of neighbors within “eps” radius.

Any point x in the dataset, with a neighbor count greater than or equal to **MinPts**, is marked as a **core point**. We say that x is **border point**, if the number of its neighbors is less than MinPts, but it belongs to the \(\epsilon\)-neighborhood of some core point z. Finally, if a point is neither a core nor a border point, then it is called a noise point or an outlier.

The figure below shows the different types of points (core, border and outlier points) using **MinPts = 6**. Here x is a core point because \(neighbours\_\epsilon(x) = 6\), y is a border point because \(neighbours\_\epsilon(y) < MinPts\), but it belongs to the \(\epsilon\)-neighborhood of the core point x. Finally, z is a noise point.



We define 3 terms, required for understanding the **DBSCAN algorithm**:

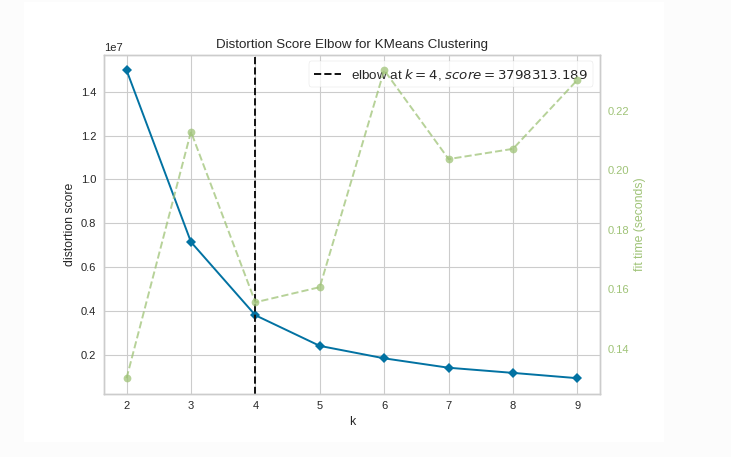
* **Direct density reachable**: A point “A” is **directly density reachable** from another point “B” if: i) “A” is in the \(\epsilon\)-neighborhood of “B” and ii) “B” is a core point.
* **Density reachable**: A point “A” is **density reachable** from “B” if there are a set of core points leading from “B” to “A.
* **Density connected**: Two points “A” and “B” are **density connected** if there are a core point “C”, such that both “A” and “B” are **density reachable** from “C”

3) What is elbow method in K means Clustering & what is it used for?

The K-Elbow Visualize implements the “elbow” method of selecting the optimal number of clusters for K-means clustering. K-means is a simple unsupervised machine learning algorithm that groups data into a specified number (k) of clusters.

we can see that the optimal number of clusters should be around 4. But visualizing the data alone cannot always give the right answer. Hence we demonstrate the following steps.  
We now define the following:-

1. **Distortion:** It is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used.
2. **Inertia:** It is the sum of squared distances of samples to their closest cluster center.



4) Discuss any 2 applications of Clustering?

1. k-Means Clustering:

k-Means is one of the most widely used and perhaps the simplest unsupervised algorithms to solve the clustering problems. Using this algorithm, we classify a given data set through a certain number of predetermined clusters or “k” clusters. Each cluster is assigned a designated cluster center and they are placed as much as possible far away from each other. Subsequently, each point belonging gets associated with it to the nearest centroid till no point is left unassigned. Once it is done, the centers are re-calculated and the above steps are repeated. The algorithm converges at a point where the centroids.

cannot move any further. This algorithm targets to minimize an objective function called the squared error function F(V) :

https://www.analytixlabs.co.in/blog/wp-content/uploads/2020/07/image.png

where, ||xi – vj|| is the distance between Xi and Vj.

Ci is the count of data in cluster C is the number of cluster centroids.

**Implementation:**

In R, there is a built-in function kmeans() and in Python, we make use of scikit-learn cluster module which has the KMeans function. (sklearn.cluster.KMeans)

**Advantages:**

1. Can be applied to any form of data – as long as the data has numerical (continuous) entities.  
2. Much faster than other algorithms.  
3. Easy to understand and interpret.

**Drawbacks:**

1. Fails for non-linear data.  
2. It requires us to decide on the number of clusters before we start the algorithm – where the user needs to use additional mathematical methods and also heuristic knowledge to verify the correct number of centers.  
3. This cannot work for Categorical data.  
4. Cannot handle outliers.

**Application Areas:**

a. Document clustering – high application area in Segmenting text-matrix related like data like DTM, TF-IDF etc.  
b. Banking and Insurance fraud detection where majority of the columns represent a financial figure – continuous data.  
c. Image segmentation.  
d. Customer Segmentation.

2. Hierarchical Clustering Algorithm

Hierarchical clustering methods follow two approaches – Divisive and Agglomerative types. Their implementation family contains two algorithms respectively, the divisive DIANA (Divisive Analysis) and AGNES (Agglomerative Nesting) for each of the approaches.

**DIANA or Divisive Analysis**

As discussed earlier, the divisive approach begins with one single cluster where all the data points belong to. Then it is split into multiple clusters and the data points get reassigned to each of the clusters on the basis of the nearest distance measure of the pair wise distance between the data points. These distance measures can be Ward’s Distance, Centroid Distance, average linkage, complete linkage or single linkage. Ideally, the algorithm continues until each data has its own cluster.

**Implementation:**

In R, we make use of the diana() function from cluster package (cluster::diana)

#### Agglomerative Nesting or AGNES

AGNES starts by considering the fact that each data point has its own cluster, i.e., if there are n data rows, then the algorithm begins with n clusters initially. Then, iteratively, clusters that are most similar – again based on the distances as measured in DIANA – are now combined to form a larger cluster. The iterations are performed until we are left with one huge cluster that contains all the data-points.

**Implementation:**

In R, we make use of the agnes() function from cluster package (cluster::agnes()) or the built-in hclust() function from the native stats package. In python, the implementation can be found in scikit-learn package via the AgglomerativeClustering function inside the cluster module(sklearn.cluster.AgglomerativeClustering)

**Advantages:**

1. No prior knowledge about the number of clusters is needed, although the user needs to define a threshold for divisions.   
2. Easy to implement across various forms of data and known to provide robust results for data generated via various sources. Hence it has a wide application area.

**Disadvantages:**

1. The cluster division (DIANA) or combination (AGNES) is really strict and once performed, it cannot be undone and re-assigned in sub sequent iterations or re-runs.  
2. It has a high time complexity, in the order of O(n^2 log n) for all the n data-points, hence cannot be used for larger datasets.  
3. Cannot handle outliers and noise

**Application areas:**

1. Widely used in DNA sequencing to analyze the evolutionary history and the relationships among biological entities (Phylogenetics).

5) Which of the following is required by K-means clustering?

a) defined distance metric

b) number of clusters

c) initial guess as to cluster centroids

d) all of the mentioned

Answer : D

6) Point out the wrong statement.

a) k-means clustering is a method of vector quantization

b) k-means clustering aims to partition n observations into k clusters

c) k-nearest neighbor is same as k-means

d) none of the mentioned

Answer C

7) Which of the following function is used for k-means clustering?

a) k-means

b) k-mean

c) heatmap

d) none of the mentioned

Answer A:

8) Clustering is a-

a) Supervised learning

b) Un Supervised learning

c) Reinforcement learning

d) None

Answer : b

9) For Clustering we don’t require

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

Answer : a

10) Which of the step is not required for K-means clustering?

a) Distance Metrics

b) Initial no of clusters

c)Initial guess as to no of centroids

d) None

Answer : D Explanation: A, B and C approach.