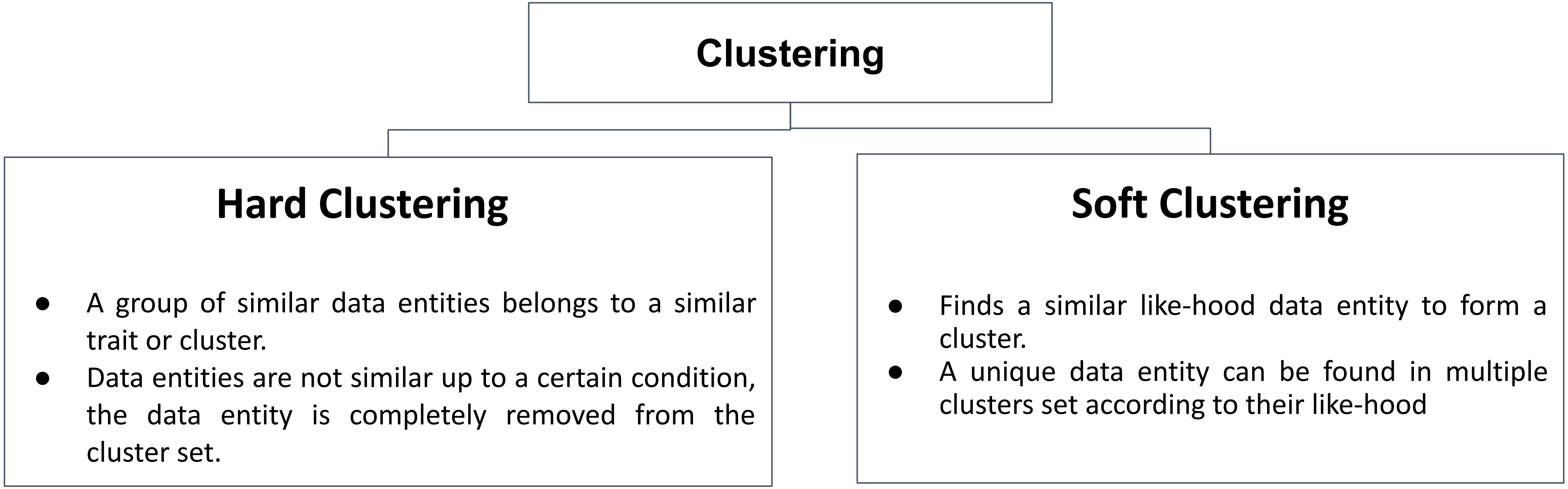
Clustering

Clustering Algorithm

- A clustering algorithm is a form of machine learning method that may be used to separate data sets depending on distinct groupings and business requirements.
- It is a well-known type of machine learning algorithm used in data science and artificial intelligence (AI).
- They are two types of clustering algorithms based on the logical grouping pattern.



Clustering

Hard Clustering

- A group of similar data entities belongs to a similar trait or cluster.
- Data entities are not similar up to a certain condition, the data entity is completely removed from the cluster set.

Soft Clustering

- Finds a similar like-hood data entity to form a cluster.
- A unique data entity can be found in multiple clusters set according to their like-hood

Clustering Methodology

Method	Explanation
Connectivity	This algorithm find the nearest similar data entity in the group of set data entities based on the notion that the data points are closer in data space. So the data entity nearer to the similar data entity will exhibit more similarity than the data entity lying very far away.
Centroid	In this type of iterative algorithm, a certain centroid point is taken into consideration first, then the similar data entity according to their closeness relative to this centroid point is set into a cluster.
Distribution	In this type of algorithm, the method finds that how much is it possible that each data entity in a cluster belongs to identical or same distribution like Gaussian or normal. One drawback of this type of algorithm is that the data set entity has to suffer from overfitting in this type of clustering.
Density	Using this algorithm, the data set is isolated with respect to different density regions of data in the data space, and then the data entity is assigned with specific clusters.
K-means	This type of clustering is used to find a local maximum after each iteration in the set of multiple data entity sets
Hierarchical	It is an algorithm which builds a hierarchy of clusters. Although, it starts with all the data points that are assigned to a cluster of their own. Then the two nearest clusters will merge into the same cluster. In the end, we use to terminate it when there is only a single cluster left.

Applications

- Anomaly detection
- Image segmentation
- Medical imaging
- Search result grouping
- Social network analysis
- Market Segmentation
- Recommendation engine

Important Questions

1) What is good clustering?

A good clustering method will produce high quality clusters in which: – the intra-class (that is, intra intra-cluster) similarity is high, the inter-class similarity is low. The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.

2) When to use K means clustering?

The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

3)What is silhouette Score?

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. It is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b

- 1: Means clusters are well apart from each other and clearly distinguished.
- 0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- -1: Means clusters are assigned in the wrong way.

4) What is meant by hierarchical Clustering

It is an algorithm which builds a hierarchy of clusters. Although, it starts with all the data points that are assigned to a cluster of their own. Then the two nearest clusters will merge into the same cluster. In the end, we use to terminate it when there is only a single cluster left.

5) Is there a way to find K value for K-means clustering?

There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

6) Explain agglomerative clustering?

The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together.

7) Why elbow method is used?

The elbow method is used **to determine the optimal number of clusters in k-means clustering**. The elbow method plots the value of the cost function produced by different values of k.

8) How the optimal k value calculated using elbow method?

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the total within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

9) What are the different methods of hierarchical clustering?

- **Agglomerative –** Also called bottom-up approach. Each observation starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy.

- **Divisive –** Also called top-down approach. All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.