ASSIGNMENT-3

MACHINE LEARNING

1. D
2. C
3. B
4. C
5. C
6. D
7. B
8. C
9. C
10. C
11. D
12. 13. Clustering is simply the grouping of data sets involving common sets of attributes and placed together in a cluster along with multiple other data sets to analyze and find inferences from it. Machine learning has two primary ‘techniques’ for creating a machine learning algorithm which are:

- Supervised learning method
- Un-supervised learning method

Clustering comes in the domain of the unsupervised learning method of machine learning, in which it draws out inferences from the data sets of variables that do not have a labeled output variable. As the name suggests, clustering is dividing the data sets into clusters such that all the different data sets present in a cluster have similar and common attributes.

It basically groups data sets with common characteristics

he entire data sets present are many for a particular problem, and it is impossible to analyze them individually; hence, clustering makes it easy to handle and gather insightful data from it. The creation of such clusters mainly depends on its creator, i.e., the programmer writing the code for it and the **algorithm** which they use.

The algorithm depends on the type of data set, the number of data sets, and the type of inferences required.

The Main Types of Clustering

Clustering has two major types:

1. Hard clustering
2. Soft clustering

To explain both, let's look at an example, assume you have to place guests in a hotel in groups, and there are 10 groups allowed to be made. Then, according to this constraint, each guest must get placed in a single group and not multiple.

In **hard clustering**, each data set must belong to a cluster completely. Considering the above example customer falls into one group out of the ten groups. Whereas **soft clustering**, a probability of a data set belonging to a cluster gets calculated, and then that data set gets placed in that cluster.

Considering the same example, you assign each guest a probability of being in one out of the ten groups.

There are also various types of clustering depending on the usage parameters, type of data set involved, and the output required. The different clustering, apart from the two general types, are:

- Constraint-based clustering
- Centroid based clustering
- Fuzzy clustering
- Hierarchal clustering
- Partition based clustering
- Grid-based clustering

Constraint-based Clustering

They base this on the approach that it can create an optimal number of data sets. The constraints defined are the required properties of the data sets and the insightful inferences to be extracted. An example of a constraint is a fixed number of clusters.

Centroid based Clustering

It is one of the simplest clustering techniques present yet. As the name suggests, in centroid-based clustering, clusters get selected as a centralized vector, and the data sets belonging close to that vector form other clusters.

Fuzzy Clustering

Fuzzy clustering breaks the commonly used barrier for clustering methods. It involves assigning a single data set to multiple clusters, and all the other clusters which are closely bound to it combine to make other clusters.

Hierarchal Clustering

Hierarchical clustering also called as Bottom-Up Approach, upholds distance metrics. In this type of clustering, each data point acts as a cluster initially, and then it groups the clusters one by one.

Partition based Clustering

This is one of the most popular clustering methods out there. In this type, clusters get divided or partitioned based upon the type of data sets involved. It helps users determine how many clusters they need to create.

Grid-based Clusters

In grid-based clustering, data sets get represented into a grid structure that comprises grids, also called cells. The overall approach of this method differs from the rest. They are more concerned with the value space surrounding the data points rather than the data points themselves.

The primary use of clustering in machine learning is to extract valuable inferences from many unstructured data sets. If you are working with large amounts of data that are also not structured, it is only logical to organize that data to make it helpful in so many other ways, and clustering helps us do that.

Clustering and classification allow you to take a sweeping glance at your data. And then form some logical structures based on what you find there before going deeper into the nuts-and-bolts analysis.

Clustering is a significant component of machine learning, and its importance is highly significant in providing better machine learning techniques.

Clustering Machine Learning Use Cases for Clustering

Clustering in machine learning has a vast range of uses and advantages over other technologies. Following are some uses of clustering in machine learning:

- Social network analysis
- Search result grouping
- Image segmentation
- Anomaly detection
- Data compression
- Privacy preservation
- Medical imaging

Takeaway

Clustering in machine learning is an essential component and makes life so much easier in creating new machine learning methods. It mainly divides many unstructured data sets into

clusters and, according to the common attributes present in them, it helps create more and more clusters.

There are two major clusters in machine learning, but multiple other algorithms and methods are also present. The uses and importance of clustering are vast and are getting more and more popular each day.

14. Clustering analysis is one of the main analytical methods in data mining. K-means is the most popular and partition based clustering algorithm. But it is computationally expensive and the quality of resulting clusters heavily depends on the selection of initial centroid and the dimension of the data. Several methods have been proposed in the literature for improving performance of the k-means clustering algorithm. Principal Component Analysis (PCA) is an important approach to unsupervised dimensionality reduction technique. This paper proposed a method to make the algorithm more effective and efficient by using PCA and modified k-means. In this paper, we have used Principal Component Analysis as a first phase to find the initial centroid for k-means and for dimension reduction and k-means method is modified by using heuristics approach to reduce the number of distance calculation to assign the data-point to cluster. By comparing the results of original and new approach, it was found that the results obtained are more effective, easy to understand and above all, the time taken to process the data was substantially reduced.