Student Name: - Kiran Digambar Yadav.

Assignment Name: - Machine Learning

Internship: - 35

Machine Learning

Question No	Answer
1	a) 2
2	d) 1,2 And 4
3	a) Interpreting and profiling clusters.
4	a) Euclidean distance.
5	b) Divisive clustering.
6	d) All answers are correct.
7	a) divide the data points into group.
8	b) Unsupervised learning.
9	a) K-means clustering.
10	a) K-means clustering.
11	d) All of the above.
12	a) Label data.

Q13 Answer: -

Clusters can be calculated using various grouping methods. This can be divided into following.

I. Graph theoretical.
II. Hierarchically.
III. Partitioning.
IV. Optimizing.

Let us understand how k-means clustering is calculated. It is a partitioning method which is particularly suitable for large amount of data.

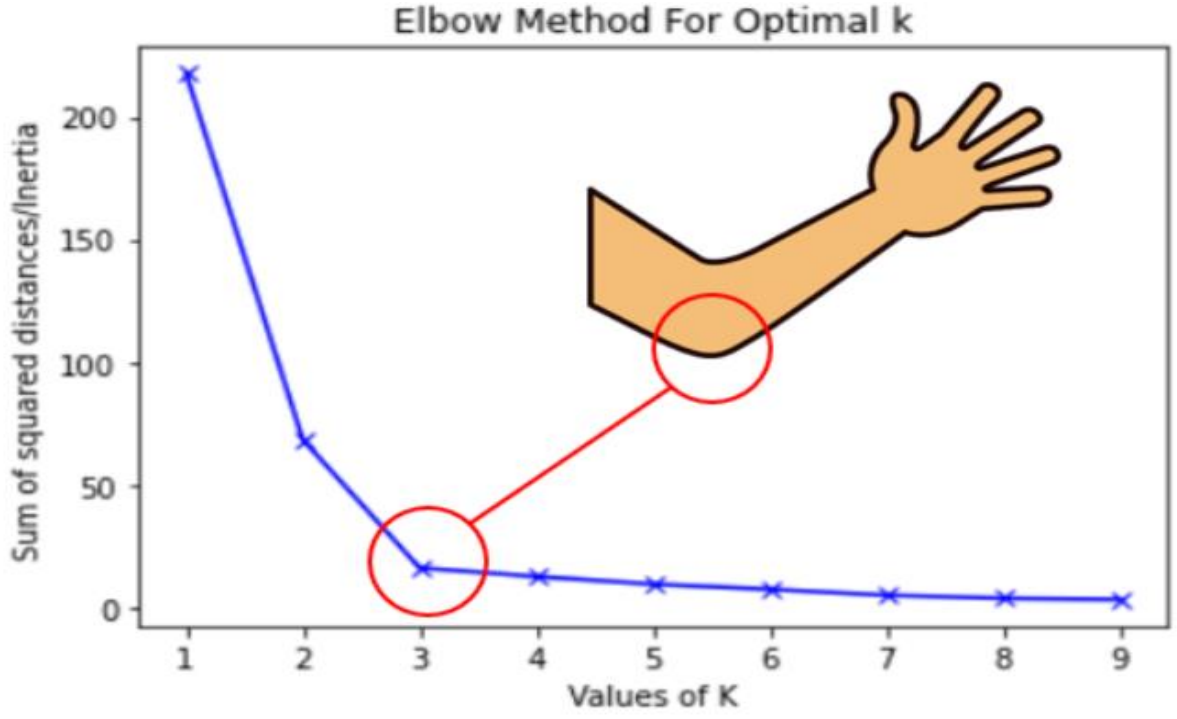
I. First an initial partition with k clusters (given number of clusters) is created.
II. Then starting with the first object in the first cluster, Euclidean distance of all objects to all clusters are calculated.
III. If an object is detected whose distance to the centre of gravity of the own cluster is greater than the distance to centre of gravity (centroid) of another cluster, this object is shifted to another cluster.
IV. Finally, the centroids of two changed clusters are calculated again, since the compositions have changed here.
V. These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (centre of cluster).

When you want to calculate clusters elbow method is used. In the elbow method we are actually varying the number of clusters (K) from 1 to 10. For each value

of K, we are calculating WCSS (Within cluster sum of square). WCSS is the sum of squared distance between each point and the centroid of clusters. When we plot the WCSS with the K value, the plot looks like an elbow. As the number of clusters increase value will start decreases. WCSS value is largest when K = 1.

When we analyse the graph, we can see that the graph is rapidly changes at a point and thus creating an elbow shape. From this point the graph starts to move almost parallel to X-axis. The K value corresponding to this point is the optimal K value or optimal number of clusters. Here we have 3 clusters. (In below image).

The dar



Line plot between K and inertia

Q14 Answer: -

If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of clustering by using the Dissimilarity/ Similarity metric in most situations. But there are some other methods to measure the qualities of good clustering, which is as below.

I. Dissimilarity/ Similarity Metric: -The similarity between the clusters can be expressed in terms of distance function, which is represented as d (i, j). distance functions are different for various data types and variables. Distance function measure is different for continuous valued variables. Distance function can be expressed as Euclidean distance, mahalanobis distance and cosine distance for different types of data.

II. Cluster completeness: - Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics, then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category. Let us consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, Q (C2, Cg) > Q (C1, Cg).

III. Ragbag: - In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category. Let us consider a clustering C1 and a cluster $C \in C1$ so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth. Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, Q (C2, Cg)>Q (C1, Cg).

IV. Small cluster preservation: - If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive. Suppose clustering C1 has split into three clusters, C11 = {d1, . . ., dn}, C12 = {dn+1}, and C13 = {dn+2}. Let clustering C2 also split into three clusters, namely C1 = {d1, . . ., dn−1}, C2 = {dn}, and C3 = {dn+1, dn+2}. As C1 splits the small category of objects and C2 splits the big category which is preferred according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, Q (C2, Cg) > Q (C1, Cg).

Q15 Answer: -

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

- Partitioning approach: - The partitioning approach constructs various partitions and then evaluates them by some criterion, e.g., minimizing the sum of square errors. It adopts exclusive cluster separation (each object belongs to exactly one group) and uses iterative relocation techniques to improve the partitioning by moving objects from one group to another. It uses a greedy approach and approach at a local optimum. It finds clusters with spherical shapes in small to medium size databases.
  - k-means
  - k-medoids
  - CLARINS
- Density-Based Method: -The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously as long as the density in the neighbourhood exceeds some threshold, i.e., for each data point within a given cluster. The radius of a given cluster has to contain at least a minimum number of points.

Density-based methods:

- DBSACN
- OPTICS

- Grid-Based Method: - In the Grid-Based method a grid is formed using the object together, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

Grid-based approach methods:

- STING
- Wave Cluster
- CLIQUE

- Model-Based Method: - In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based

on standard statistics, taking outlier or noise into account. Therefore, it yields robust clustering methods.

- Constraint-Based Method: - The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.
- Hierarchical Method: - In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:
  - Agglomerative Approach: -The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.
  - Divisive Approach: - The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.