Q&A

08 March 2020 06:54

1.) Clustering
Clustering is an unsupervised machine learning technique. It is used to place the data elements into related groups without any prior knowledge of the group definitions. Select which of the following is a clustering task?
A baby is given some toys to play. These toys consist of various animals, vehicles and houses, but the baby is unaware of these categories. The baby chooses different toys and starts making different groups with the toys based on what he feels are similar toys.
Feedback :*Since the baby is classifying objects based on its features and has no idea of any pre existing classification, it is a clustering task.*

2.) Clustering
We have learnt about three different types of machine learning techniques - regression, classification and clustering. Which of the following techniques does not require bifurcation of data points into dependent and independent variables?
Clustering
Feedback :*In clustering, we group the data points into different categories based on the given set of attributes. There are no dependent and independent variables*

3.) Type of Segmentation
A telecom company classifies its prepaid mobile customers into three types mainly based on the number of times they recharge per month. This is a type of:
Behavioral segmentation
Feedback :*Recharge is a behaviour which can be observed, opposed to attitude which resides in the mindset of the customer.*

4.) Clustering
An international foods and beverages company wants to look at what products it should launch in India. For that, it has first tried to segment the market. It is known that people living in the same area and having similar salaries will have similar eating habits. Then which of the following can be segmented in 1 group? (All options in 1 bracket are 1 segment)
A. High Earning Individual from Bengaluru
B. Low Earning Individual from Rural Uttar Pradesh
C. Mid Earning Individual from Mumbai
D. High Earning Individual from Hyderabad
(A,D) - (B) - (C)
Feedback :*You want to ensure that the people in One segment are very similar and the people in different segments are very dissimilar. So Option C is good segment to make*

5.) Segmentation Factors
You learnt that clustering is commonly used for segmenting customers. Can you think of some features on which you would want to segment the customers of an online store? Go back and check the data 'online retail' if needed.
Suggested Answer
Some of the features can be how much the people spend on buying goods from the store. Another one can be the number of times they bought goods in the last one year. Finally, you can also look at the time they last bought a good from the store.

6.) Segmentation Types
You are an analyst at a global laptop manufacturer and are given the task of deciding whether the company should enter the Indian Market. You try to estimate the market size by first breaking the market by different types of people who use a laptop such as students, working professionals and their paying capacity to get an estimate of the total market size and the characteristics of each segment. In essence, you are doing:

Demographic Segmentation
Feedback :*You are doing a demographic segmentation, since you are looking at the income and the profession of people. Notice how this is much simpler than finding data about actual laptop purchasing history of customers and then trying to estimate the market size based on that.*

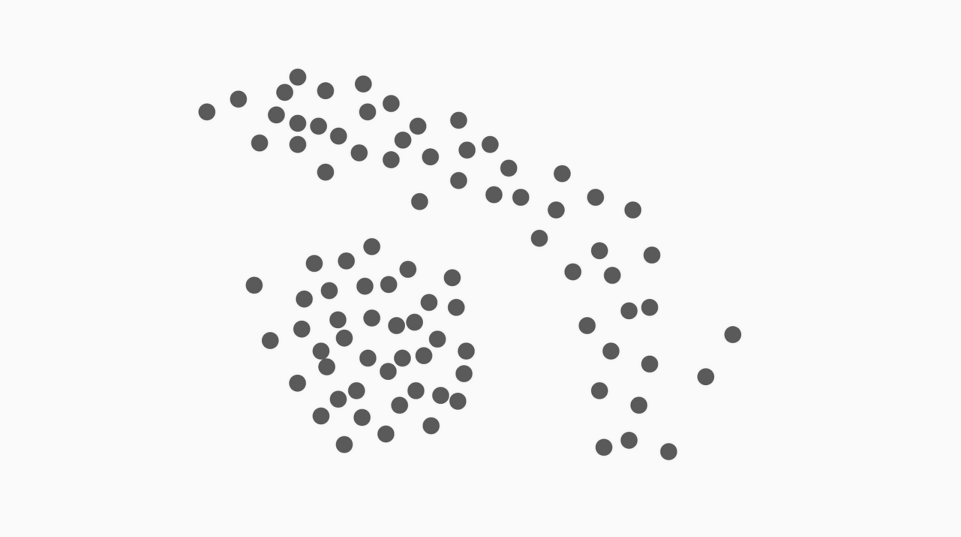
7.) K-Means algorithm
Find the distance of each of the 10 points from the two cluster centers (in column H & I) and fill the cells S6:T16. Select the distance formula to be used to find the distance of a point (xi,yi) from a centre (Xi,Yi).
SQRT(((Xi-xi)^2) + ((Yi-yi)^2))
Feedback :
This is the formula for the Euclidean distance. You can see that once the cells S6:T16 are filled with the required distances, the points are assigned to one of the clusters (marked in column E) based on the minimum distance.

8.) K-Means algorithm
In the next step of k-means clustering, you need to find the new cluster centers (H22:I23). Select the correct method used to find the new cluster centers.
Calculate the centroid of the points assigned to a particular cluster in the previous step
Feedback :*The way is new center is calculated in K-Mean clustering is the mean of all the data points belonging to that cluster.Notice that the new centers are already filled in the excel sheet.*

9.) K-Means algorithm
Repeating the previous two steps again, you get the new cluster centers (H38:I39).Continue this process until the algorithm converges (Two consecutive iterations have the same center). What are the x and y coordinates of the center of cluster 1 finally?
5.6, 5.4
Feedback :*Notice the clusters formed on the chart with their centers. You can explore k the -means algorithm by randomly assigning the centers and looking at the number of iterations needed for convergence.*

10.) K-Means
What is the significance of "argmin" in the assignment step equation?
Suggested Answer
For a ith data point which is a 2d object and μ which is again a 2d object, we compute the distance between these two, this is given by d(xi,μk) where k is the number of clusters and then from these k different results we will choose the minimum of all.

Now look at the given image and answer the question that follows:



11.) K-Means algorithm

Consider the above arrangement of points. How many clusters do you intuitively feel are present. What will happen if you use K-Means clustering here? How do you think this problem can be solved?
Suggested Answer
Intuitively, it looks like 2 clusters are present. If we use K means, then we will get wrong clusters since the points in the outer ring like structure will not be segmented accurately. A reason for that is K-Means looks for how close the points are to a centroid and this distance or measure of closeness is the "linear distance". One way to correct this can be to see the distance between all the points and then cluster the closest points.

12.) K-Means algorithm
In this exercise, you will perform k-means clustering manually on a small dataset. Consider the following dataset having 2 features and 6 observations.

Observation number	**X1**	**X2**
1	**1**	**4**
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

Use k = 2 for the entire exercise.
Assign clusters to each observation such that the odd numbered observations get assigned cluster = 1, i.e. points 1, 3 and 5 get assigned cluster = 1 and the even ones get assigned cluster = 2.
Compute the centroid of the two clusters
Assign each observation to the centroid to which it is closest (using euclidean distance). Report the new cluster labels for each observation.
Repeat the above steps until the clusters stop changing

The observations (identified by their row numbers) in the two clusters respectively are:
(1, 2, 3) and (4, 5, 6)
Feedback :*You will see that the solution converges really fast. After just the first iteration, the solution converges. After the first iteration, observation 1,2,3 get assigned to cluster 1 and observation 4,5,6 get assigned to cluster 2. The initial cluster centroids were (2.3,3.3) and (3.3,1.3) and the distances from these is calculated to assign the clusters.*

13.) K-Means algorithm
Did you take a note of one of the most important underlying principles of k-means algorithm. Select the option which describes the principle correctly.
Maximise the inter-cluster distance and minimise the intra-cluster distance
Feedback :*K means clustering tries to minimise the intra cluster distance and maximise the inter cluster distance.*

14.) K-Means algorithm
If we are worried about K-means getting stuck in bad local optima, one way to solve this problem is if we try using multiple random initializations. Is this true or false? You can read about local optimum and global optimum here
True
Feedback :*Since each run of K-means is independent, multiple runs can find different local optima, and this can help in choosing the global optimum value.*

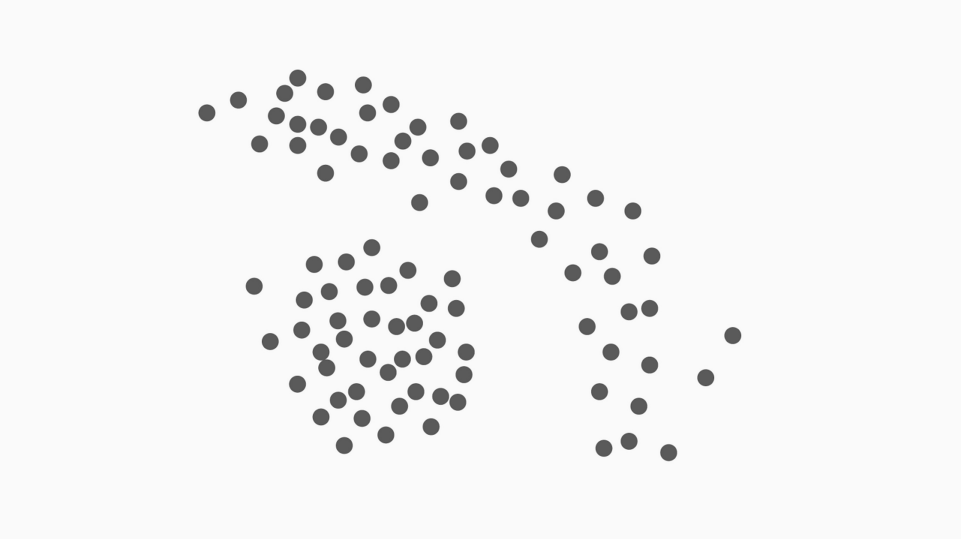
15.) Using K-Means Clustering
What do you think will happen if you run the clustering without scaling the data?
Suggested Answer
Since the scales of the data are different, more weightage will be given to Monetary value. Data points which have very different monetary values will be classified differently, even though they

might actually be very similar in Recency and Frequency.

16.) K - Means in Python
Which parameters do you think are the most important for segmenting the states? How did you decide this?
Suggested Answer
The parameters used depend on the question at hand. We can choose different age groups and different categories such as graduate or above etc.

17.) K - Means in Python
How will you check if the segmenting is good or whether you need to use different factors for segmenting?
Suggested Answer
One really easy way is to see if the .are logically correct. For example, we can check if states in a similar geography and economic situation are clustered together or not. You can also run hypothesis test to check whether the population of different clusters is significantly different or not, If you look at the data, you can see that some specific customers or some specific states should be grouped together.

18.) K - Means in Python
How are the clusters different when we have not scaled compared to clusters formed after scaling?
Illiteracy percentage gets higher weightage when there is no scaling
Feedback :
Look at the clusters formed with and without scaling. You will see that for the ones formed without scaling, the states with similar literacy rates will fall in the same cluster, even though their graduate percentage differs.



19.) Hierarchical Clustering
You had made clusters for it using the K-Means algorithm. How do you think clusters will be made using hierarchical algorithm on this data?
Suggested Answer
Since now you are looking at the closest distance between clusters, you will get two clusters - one at the center and the one which contains the points at the edges.

20.)Hierarchical Clustering
Look at the following matrix. This is the distance matrix between 4 points - A, B,C, D. Find out which 2 clusters will merge first.

	A	B	C	D
A				
B	2.24			

C	8.06	10.00		
D	5.83	8.06	5.00	

A-B
Feedback :*Look at the data and see that the minimum distance is between A and B*

Comprehension - Hierarchical Clustering Algorithm
Given below are five data points having two attributes x and y:

Observation	**x**	**y**
1	3	2
2	3	5
3	5	3
4	6	4
5	6	7

The distance matrix of the points, indicating the Euclidean distance between points, is as follows:

Label	1	2	3	4	5
1	0.00	3.00	2.24	3.61	5.83
2	3.00	0.00	2.83	3.16	3.61
3	2.24	2.83	0.00	1.41	4.12
4	3.61	3.16	1.41	0.00	3.00
5	5.83	3.61	4.12	3.00	0.00

Take the distance between two clusters as the minimum distance between the points in the two clusters. Based on this information, answer the following questions.

21.) Hierarchical Clustering
How many clusters are there initially (before any fusions have happened)?
5
Feedback :*Since this is agglomerative clustering, initially, all the points are 1 cluster.*

22.) Hierarchical Clustering
Which two clusters will be fused first?
3 and 4
Feedback :*These two points(clusters) have the minimum distance.*

23.) Hierarchical Clustering
Which clusters will be fused in step two?
1 will be fused with the cluster(3, 4)
Feedback :*The distance between points 1 and 3 is 2.24 units, which is the minimum among all the new clusters. Hence they will be joined now.*

24.) Hierarchical Clustering
How many total clusters are there right after point number 1 fuses with the cluster(3, 4)?
3
Feedback :*Since three points have fused into 1 cluster, total clusters left are (1,3,4) - (2) - (5).*

25.) Hierarchical Clustering
Which clusters will be fused after 1 fuses with (3, 4)?
2 will fuse with the cluster (1, 3, 4)

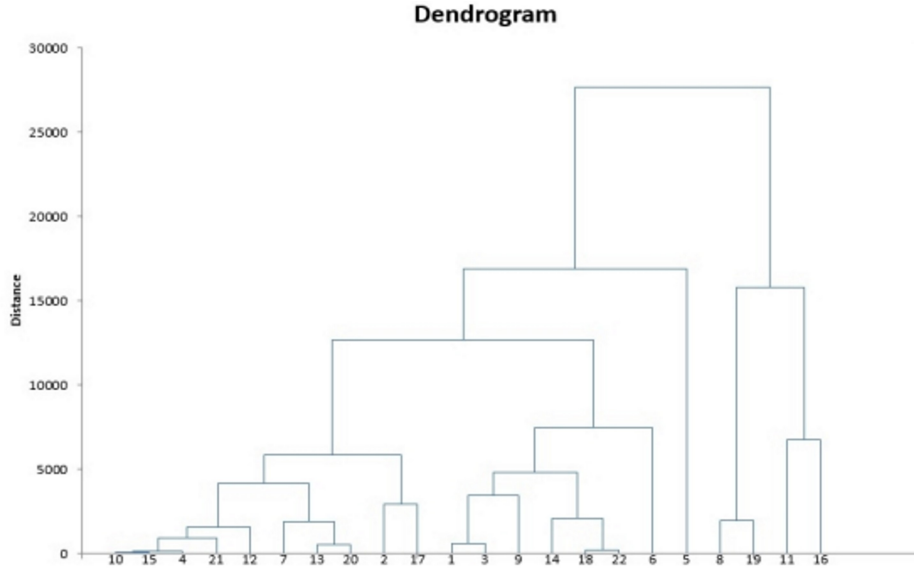
Feedback :*The distance of point 2 from point 3 is 2.83 units, whereas the minimum distance of point 5 from the cluster (1,3,4) is 3 units.*

26.) Hierarchical Clustering
What happens in the last step of the algorithm?
5 fuses with (1, 2, 3, 4)
Feedback :*Since all the other points are already part of one cluster, the last point will also join that cluster at this step.*

27.) Hierarchical Clustering
Select the appropriate option which describes the Complete Linkage method.
Incorrect
In complete linkage hierarchical clustering, the inter cluster distance is defined as the longest distance between two points (one point in each cluster)
Feedback :*In the complete linkage, inter cluster distance is calculated as the maximum distance between 2 points (one in each cluster), However, the point is assigned to a new cluster basis it's minimum distance from the clusters*

28.) Hierarchical Clustering
How many iterations are required to form the final single cluster?
5
Feedback :*Initially, n clusters are made and in each iteration, the number of clusters gets reduced by 1. So the number of iterations required is n-1. Here n = number of points = 6. So the correct answer is 5.*

Now look at the following dendrogram and answer the questions that follow.



Hierarchical Clustering

29.) Hierarchical Clustering
Consider the above dendrogram for agglomerative clustering and answer the following questions.
Find number of clusters if threshold value is 10000. (refer fig)
5
Feedback :*Draw a horizontal line at that height. It cuts 5 vertical lines, all of which represent a cluster.*

30.) Hierarchical Clustering
Find the threshold value if the no. of clusters to be formed is 4. (refer fig)
15000
Feedback :*Look at the height at which a horizontal line will cut 4 vertical lines.*

31.) Hierarchical vs K-Means

What are the benefits of Hierarchical Clustering over K-Means clustering? What are the disadvantages?
Suggested Answer
Hierarchical clustering generally produces better clusters, but is more computationallyintensive.

32.) Hierarchical Clustering
Can you use the dendrogram to make meaningful clusters? (By looking at which elements leave and join at what height)
Suggested Answer
Yes. It is a great tool. You can look at what stage an element is joining a cluster and hence see how similar or dissimilar it is to the rest of the cluster. If it joins at the higher height, it is quite different from the rest of the group. You can also see which elements are joining which cluster at what stage and can thus use business understanding to cut the dendrogram more accurately.

33.) Hierarchical Clustering
Compare the different linkages. Which one do you think gives a well-separated dendrogram? Are there any advantages of that?
Suggested Answer
Average and Complete linkage methods give a well-separated dendrogram, whereas single linkage gives us dendrograms which are not very well separated. We generally want well separated clusters.

34.) DBSCAN is a density-based clustering algorithm that divides a data set into subgroups of high-density regions. DBSCAN groups together point that are close to each other based on a distance measurement (usually Euclidean distance) and a minimum number of points. It also marks as outliers the points that are in low-density regions.
DBSCAN algorithm requires 2 parameters i.e. Epsom or EPS and MinPoints or MinSamples.