1. A set of one-dimensional data points is given to you: 5, 10, 15, 20, 25, 30, 35. Assume that k = 2 and that the first set of random centroid is 15, 32, and that the second set is 12, 30. ?

Using the k-means method, create two clusters for each set of centroid described above.

For each set of centroid values, calculate the SSE.

2. Describe how the Market Basket Research makes use of association analysis concepts ?

Ans : Market Basket Analysis is one of the key techniques used by large retailers to uncover associations between items. It works by looking for combinations of items that occur together frequently in transactions. To put it another way, it allows retailers to identify relationships between the items that people buy.

In market basket analysis (also called association analysis or frequent itemset mining), you analyze purchases that commonly happen together. For example, people who buy bread and peanut butter also buy jelly. Or people who buy shampoo might also buy conditioner.market basket analysis, association rules are used to predict the likelihood of products being purchased together. Association rules count the frequency of items that occur together, seeking to find associations that occur far more often than expected.

3. Give an example of the Apriori algorithm for learning association rules ?

Ans : Apriori Algorithm is one of the algorithm used for transaction data in Association Rule Learning. It allows us to mine the frequent itemset in order to generate association rule between them. Example: list of items purchased by customers, details of website which are frequently visited etc.Association Rules. The Apriori algorithm calculates rules that express probabilistic relationships between items in frequent itemsets For example, a rule derived from frequent itemsets containing A, B, and C might state that if A and B are included in a transaction, then C is likely to also be included.classic example of association rule mining refers to a relationship between diapers and beers. The example, which seems to be fictional, claims that men who go to a store to buy diapers are also likely to buy beer. Data that would point to that might look like this: A supermarket has 200,000 customer transactions.

4. In hierarchical clustering, how is the distance between clusters measured? Explain how this metric is used to decide when to end the iteration ?

Ans : In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.Euclidean distance

For most common clustering software, the default distance measure is the Euclidean distance. Depending on the type of the data and the researcher questions, other dissimilarity measures might be preferred.

5. In the k-means algorithm, how do you recompute the cluster centroids ?

Ans : To calculate the centroid from the cluster table just get the position of all points of a single cluster, sum them up and divide by the number of points.k

means is one of the most popular clustering algorithms. K-means groups observations by minimizing distances between them and maximizing group distances. One of the primordial steps in this algorithm is centroid selection, in which k initial centroids are estimated either randomly, calculated, or given by the user.centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

6. At the start of the clustering exercise, discuss one method for determining the required number of clusters ?

Ans : Elbow method.

2 X-means clustering.

3 Information criterion approach.

4 Information–theoretic approach.

5 Silhouette method.

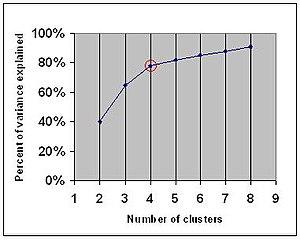
6 Cross-validation.

7 Finding number of clusters in text databases.

8 Analyzing the kernel matrix.

simple method to calculate the number of clusters is to set the value to about √(n/2) for a dataset of 'n' points.

Elbow method



The elbow method looks at the percentage of explained variance as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified,[1] making this method very subjective and unreliable. Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance.

7. Discuss the k-means algorithm's advantages and disadvantages ?

Ans : Below are the advantages mentioned:

It is fast

Robust

Easy to understand

Comparatively efficient

If data sets are distinct, then gives the best results

Produce tighter clusters

When centroids are recomputed, the cluster changes.

Flexible

Easy to interpret

Better computational cost

Enhances Accuracy

Works better with spherical clusters

Disadvantage

Needs prior specification for the number of cluster centers

If there are two highly overlapping data, then it cannot be distinguished and cannot tell that there are two clusters

With the different representations of the data, the results achieved are also different

Euclidean distance can unequally weigh the factors

It gives the local optima of the squared error function

Sometimes choosing the centroids randomly cannot give fruitful results

It can be used only if the meaning is defined

Cannot handle outliers and noisy data

Do not work for the non-linear data set

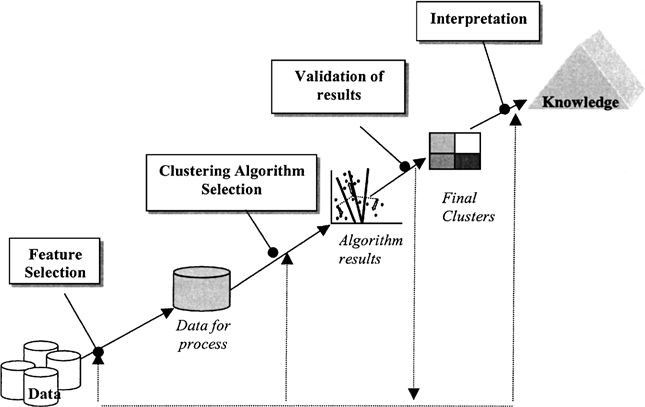
Lacks consistency

Sensitive to scale

If very large data sets are encountered, then the computer may crash.

Prediction issues

8. Draw a diagram to demonstrate the principle of clustering ?



9. During your study, you discovered seven findings, which are listed in the data points below. Using the K-means algorithm, you want to build three clusters from these observations. The clusters C1, C2, and C3 have the following findings after the first iteration ?

C1: (2,2), (4,4), (6,6); C2: (2,2), (4,4), (6,6); C3: (2,2), (4,4),

C2: (0,4), (4,0), (0,4), (0,4), (0,4), (0,4), (0,4), (0,4), (0,

C3: (5,5) and (9,9)

What would the cluster centroids be if you were to run a second iteration? What would this clustering's SSE be?

Ans : Finding centroid for data points in cluster C1 = ((2+4+6)/3, (2+4+6)/3) = (4, 4)

Finding centroid for data points in cluster C2 = ((0+4)/2, (4+0)/2) = (2, 2)

Finding centroid for data points in cluster C3 = ((5+9)/2, (5+9)/2) = (7, 7)

Hence, C1: (4,4), C2: (2,2), C3: (7,7)

10. In a software project, the team is attempting to determine if software flaws discovered during testing are identical. Based on the text analytics of the defect details, they decided to build 5 clusters of related defects. Any new defect formed after the 5 clusters of defects have been identified must be listed as one of the forms identified by clustering. A simple diagram can be used to explain this process. Assume you have 20 defect data points that are clustered into 5 clusters and you used the k-means algorithm ?