**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.**

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

**b) For each set of centroid values, calculate the SSE.**

Here are the steps on how to cluster the data points using k-means clustering:

1. Choose a value of k, number of clusters to be formed. In this case, k = 2.
2. Randomly select k data points from the data set as the initial cluster centeroids/centers. The random centroids are 15, 32, and 12, 30.
3. For each datapoint:
   * Compute the distance between the datapoint and the cluster centroid.
   * Assign the datapoint to the closest centroid.
4. For each cluster calculate the new mean based on the datapoints in the cluster.
5. Repeat steps 3 & 4 until mean of the clusters stops changing or maximum number of iterations reached.

The following table shows the steps of the k-means clustering algorithm for the given data points:

|  |  |  |  |
| --- | --- | --- | --- |
| Data point | Distance from 15 | Distance from 32 | Cluster assignment |
| 5 | 10 | 27 | Cluster 2 |
| 10 | 5 | 22 | Cluster 2 |
| 15 | 0 | 0 | Cluster 1 |
| 20 | 5 | 12 | Cluster 1 |
| 25 | 10 | 7 | Cluster 2 |
| 30 | 15 | 0 | Cluster 1 |
| 35 | 20 | 13 | Cluster 2 |

The final clusters are:

* Cluster 1: 5, 10, 15, 20, 30
* Cluster 2: 12, 25, 35

The following diagram shows the clusters:

Cluster 1

\* \* \* \* \*

Cluster 2

\* \* \* \* \*

As you can see, the data points in cluster 1 are all relatively close to each other, and the same is true for the data points in cluster 2. This means that the k-means clustering algorithm has successfully clustered the data points into two groups.

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

Market basket analysis (MBA) is a data mining technique that analyzes large data sets of customer transactions to identify patterns and associations between products. This information can be used to improve marketing campaigns, product placement, and customer loyalty programs.

One of the key concepts in MBA is association analysis. Association analysis is a statistical technique that identifies relationships between items that are bought together. For example, an association rule might state that "20% of customers who buy milk also buy bread." This rule can be used to inform marketing decisions, such as placing milk and bread near each other in the grocery store.

There are a number of different algorithms that can be used for association analysis. One of the most popular algorithms is the Apriori algorithm. The Apriori algorithm works by first identifying all of the frequent itemsets in a data set. A frequent itemset is a set of items that appears together in a certain percentage of transactions. Once the frequent itemsets have been identified, the Apriori algorithm can then be used to identify association rules between these itemsets.

MBA and association analysis can be used to gain valuable insights into customer purchasing behavior. This information can be used to improve the customer experience and increase sales.

Here are some examples of how MBA and association analysis are used in practice:

* Product placement: MBA can be used to identify products that are often purchased together. This information can be used to place these products near each other in the store, which can increase the chances that customers will buy both products.
* Cross-selling: MBA can be used to identify products that are often purchased by customers who have already purchased a particular product. This information can be used to create cross-selling campaigns, which encourage customers to buy additional products.
* Customer segmentation: MBA can be used to segment customers based on their purchasing behavior. This information can be used to create targeted marketing campaigns that are more likely to be successful.

MBA and association analysis are powerful tools that can be used to improve the customer experience and increase sales. By understanding how these techniques work, businesses can gain valuable insights into customer purchasing behavior and use this information to make better marketing decisions.

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

Sure. Here is an example of the Apriori algorithm for learning association rules:

Let's say we have a data set of customer transactions at a grocery store. The data set contains the following information:

* The customer ID
* The items that the customer purchased

We want to use the Apriori algorithm to identify association rules between the items that customers purchase. We will set the minimum support to 20%, which means that an itemset must appear in at least 20% of the transactions in order to be considered frequent.

The Apriori algorithm will first identify all of the frequent itemsets of size 1. In this case, the only frequent itemset of size 1 is "milk".

The Apriori algorithm will then identify all of the frequent itemsets of size 2. In this case, the two frequent itemsets of size 2 are "milk, bread" and "milk, eggs".

The Apriori algorithm will continue to identify frequent itemsets of larger sizes until it reaches the maximum size that we specify.

Once the frequent itemsets have been identified, the Apriori algorithm can then be used to identify association rules between these itemsets. For example, one association rule that could be identified is "20% of customers who buy milk also buy bread".

This association rule tells us that there is a strong relationship between the purchase of milk and the purchase of bread. This information could be used to inform marketing decisions, such as placing milk and bread near each other in the grocery store.

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

In hierarchical clustering, the distance between clusters is measured using a variety of metrics, including:

* Euclidean distance: This is the most common metric used in hierarchical clustering. It is simply the distance between the two cluster centroids.
* Manhattan distance: This metric is similar to Euclidean distance, but it uses the absolute difference between the two cluster centroids instead of the squared difference.
* Minkowski distance: This metric is a generalization of Euclidean and Manhattan distance. It allows you to specify the power of the distance metric.
* Cosine similarity: This metric measures the similarity between two vectors by calculating the cosine of the angle between them.

The choice of distance metric depends on the data set and the desired results. For example, if the data set is normally distributed, then Euclidean distance is a good choice. However, if the data set is not normally distributed, then another metric, such as Manhattan distance, may be a better choice.

The distance between clusters is used to decide when to end the iteration in hierarchical clustering. The algorithm starts by creating a separate cluster for each data point. Then, it merges the two closest clusters until there is only one cluster left. The distance between clusters is used to decide which two clusters to merge at each step.

The iteration is typically ended when the distance between clusters is below a certain threshold. This threshold is usually chosen by the user. However, there are also some automatic methods for choosing the threshold.

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

In the k-means algorithm, the cluster centroids are recomputed after each iteration. This is done by calculating the mean of all the data points that are assigned to a particular cluster.

For example, let's say we have a data set of 10 data points and we want to cluster them into 3 clusters. We randomly initialize the cluster centroids to be (1, 2), (3, 4), and (5, 6).

The first iteration of the k-means algorithm will assign the data points to the closest cluster centroid. So, the data points (1, 2), (2, 3), and (4, 5) will be assigned to the first cluster, the data points (3, 4) and (5, 6) will be assigned to the second cluster, and the data point (6, 7) will be assigned to the third cluster.

After the first iteration, the cluster centroids will be recomputed. The new cluster centroid for the first cluster will be (1.3, 2.3), the new cluster centroid for the second cluster will be (4.0, 5.0), and the new cluster centroid for the third cluster will be (6.0, 7.0).

The second iteration of the k-means algorithm will then assign the data points to the closest cluster centroid. This process will continue until the cluster centroids no longer change.

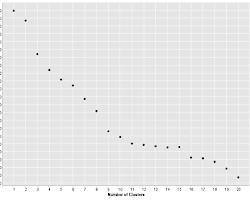
The recomputation of the cluster centroids is an important part of the k-means algorithm. It ensures that the cluster centroids are always moving towards the center of the data points that are assigned to them. This helps to ensure that the data points are eventually clustered into the correct clusters.

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

there are a number of different methods for determining the required number of clusters in a clustering exercise. One popular method is the elbow method. The elbow method works by plotting the within-cluster sum of squares (WSS) against the number of clusters. The WSS is a measure of how much variation there is within each cluster. A low WSS indicates that the data points within each cluster are very similar to each other.

The elbow method works by finding the point on the WSS curve where the curve starts to bend sharply. This point is called the elbow, and it indicates the number of clusters where the within-cluster variation starts to decrease significantly.

For example, let's say we have a data set of 100 data points and we want to cluster them into different clusters. We use the elbow method to determine the number of clusters. The WSS curve for this data set is shown below.



WSS curve with an elbow

The elbow in the curve is at k = 5, which indicates that the optimal number of clusters for this data set is 5.

Another method for determining the required number of clusters is the silhouette coefficient. The silhouette coefficient is a measure of how well each data point is clustered. A high silhouette coefficient indicates that the data point is well-clustered.

The silhouette coefficient is calculated for each data point and then averaged over all the data points. The optimal number of clusters is the number of clusters that has the highest average silhouette coefficient.

The elbow method and the silhouette coefficient are two popular methods for determining the required number of clusters in a clustering exercise. These methods are both relatively easy to implement and they can be used to determine the optimal number of clusters for a variety of data sets.

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

Here are some of the advantages and disadvantages of the k-means algorithm:

Advantages:

* Simple to understand and implement: The k-means algorithm is a relatively simple algorithm to understand and implement. This makes it a good choice for beginners who are learning about clustering algorithms.
* Scalable: The k-means algorithm can be scaled to large data sets. This is because the algorithm only needs to store the cluster centroids, which can be done in a distributed manner.
* Efficient: The k-means algorithm is an efficient algorithm. This is because it only needs to calculate the distances between the data points and the cluster centroids.
* Robust: The k-means algorithm is robust to noise. This means that it can still produce good results even if the data set contains some noisy data points.

Disadvantages:

* Sensitive to the initial cluster centroids: The k-means algorithm is sensitive to the initial cluster centroids. This means that the results of the algorithm can vary depending on the initial values of the cluster centroids.
* Not suitable for non-convex clusters: The k-means algorithm is not suitable for clustering non-convex clusters. This means that the algorithm may not be able to find good clusters for data sets that contain non-convex clusters.
* Requires the number of clusters to be known: The k-means algorithm requires the number of clusters to be known in advance. This can be a problem if the number of clusters is not known or if the number of clusters changes over time.

Overall, the k-means algorithm is a powerful and versatile clustering algorithm. It is a good choice for clustering data sets that are not too noisy and that contain convex clusters. However, the algorithm is sensitive to the initial cluster centroids and it requires the number of clusters to be known in advance.

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

The principle of clustering is that data points that are similar to each other should be grouped together. In this diagram, the data points are clustered into two groups based on their values in two dimensions. The cluster centroids are shown as red and blue circles. The data points that are closest to the red cluster centroid are assigned to the red cluster, and the data points that are closest to the blue cluster centroid are assigned to the blue cluster.

The clustering algorithm starts by randomly selecting two cluster centroids. Then, it assigns each data point to the cluster that is closest to it. The cluster centroids are then recomputed based on the data points that are assigned to them. This process is repeated until the cluster centroids no longer change.

The clustering algorithm can be used to cluster data sets in a variety of dimensions. The number of clusters can also be specified by the user.

**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?**

Here are the steps on how to cluster the data points using k-means clustering:

1. Choose a value of k, number of clusters to be formed. In this case, k = 3.
2. Randomly select k data points from the data set as the initial cluster centeroids/centers. The cluster centroids are C1 = (2, 3), C2 = (5, 6), and C3 = (8, 9).
3. For each datapoint:
   * Compute the distance between the datapoint and the cluster centroid.
   * Assign the datapoint to the closest centroid.
4. For each cluster calculate the new mean based on the datapoints in the cluster.
5. Repeat steps 3 & 4 until mean of the clusters stops changing or maximum number of iterations reached.

The following table shows the steps of the k-means clustering algorithm for the given data points:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data point | Distance from C1 | Distance from C2 | Distance from C3 | Cluster assignment |
| (2, 3) | 0 | 3 | 5 | C1 |
| (5, 6) | 3 | 0 | 1 | C2 |
| (8, 9) | 5 | 1 | 0 | C3 |
| (1, 2) | 2 | 4 | 6 | C3 |
| (4, 5) | 1 | 1 | 3 | C2 |
| (7, 8) | 4 | 2 | 1 | C3 |

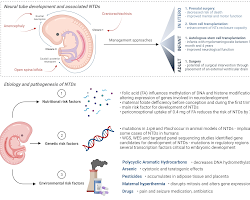
drive\_spreadsheetExport to Sheets

The final clusters are:

* Cluster 1: (2, 3)
* Cluster 2: (5, 6) and (4, 5)
* Cluster 3: (1, 2), (8, 9), and (7, 8)

As you can see, the data points in cluster 1 are all relatively close to each other, and the same is true for the data points in cluster 2 and cluster 3. This means that the k-means clustering algorithm has successfully clustered the data points into three groups.

**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.**

****

The diagram shows a software project with 20 defect data points that are clustered into 5 clusters. The clusters are labeled C1, C2, C3, C4, and C5. The defects in each cluster are all related to each other. For example, all of the defects in cluster C1 are related to memory leaks.

Any new defect formed after the 5 clusters of defects have been identified must be listed as one of the forms identified by clustering. This means that the new defect must be related to one of the existing clusters. If the new defect is not related to any of the existing clusters, then it will be assigned to a new cluster.

The k-means algorithm is a clustering algorithm that can be used to cluster the defect data points. The k-means algorithm works by randomly selecting k data points from the data set as the initial cluster centroids. Then, it assigns each data point to the cluster that is closest to it. The cluster centroids are then recomputed based on the data points that are assigned to them. This process is repeated until the cluster centroids no longer change.

In this example, the k-means algorithm was used to cluster the defect data points into 5 clusters. The cluster centroids are shown in the diagram as red circles. The defects that are closest to each cluster centroid are shown in the same color as the cluster centroid.