Assignment 19 Solutions

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. ?

1. Using the k-means method, create two clusters for each set of centroid described above.
2. For each set of centroid values, calculate the SSE.

Ans: a) Cluster 1 (using 15, 32 as centroids): 5, 10, 15, 20, 25 // Cluster 2 (using 15, 32 as centroids): 30, 35

b) For the set of centroid values 15, 32, the SSE = 15 + 25 = 40. For the set of centroid values 12, 30, the SSE = 10 + 25 = 35.

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

Ans: Market Basket Research (MBR) is a data mining technique that uses association analysis concepts to identify patterns in customer behavior. Association analysis is used to identify relationships between items that people purchase together. MBR uses this information to identify frequent co-occurrences of items and to create “rules” that can be used to predict future customer behavior. MBR can be used to indicate which items should be recommended to a customer based on their past purchases, or to identify items that should be placed together in the same section of a store. In addition, it can provide insights into customer preferences, enabling retailers to tailor their marketing efforts and product offerings accordingly.

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

Ans: Example:

Let's say we have a dataset of items purchased in a grocery store. We can use the Apriori algorithm to find association rules that describe relationships between items.

For example, we might find an association rule like:

If a customer buys milk, they are likely to also buy eggs (support = 0.7, confidence = 0.8).

This rule tells us that 70% of customers who buy milk also buy eggs and that 80% of customers who buy eggs also buy milk.

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: The distance between clusters is typically measured by calculating the Euclidean distance or the Manhattan distance between the centroids of the clusters. This metric is used to decide when to end the iteration in hierarchical clustering by setting a threshold value. Once the distance between the centroids of the clusters is less than or equal to the threshold value, the iteration is stopped and the clusters are considered to be complete.

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

Ans: To recompute the cluster centroids, the mean of all data points in each cluster needs to be calculated. This mean is then used to update the centroid position. To do this, the coordinates of each data point in the cluster are added together and then divided by the number of data points in the cluster. This new mean value is then used as the new centroid position.

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

Ans: One method for determining the required number of clusters is the elbow method. This method involves plotting the within-cluster sum of squared errors (WCSS) against the number of clusters and identifying the number of clusters where the WCSS begins to plateau. This is the point at which adding more clusters would not significantly improve the WCSS.

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

Ans: Advantages of K-Means Algorithm:

1. K-Means is an efficient and simple algorithm for clustering data.
2. K-Means is robust to outliers and noise in the data.
3. K-Means is relatively fast and computationally inexpensive.
4. K-Means produces tighter clusters than hierarchical clustering.

Disadvantages of K-Means Algorithm:

1. K-Means requires prior knowledge of the number of clusters (K) to be generated.
2. K-Means is sensitive to the initial seed points and may produce different results each time it is run.
3. K-Means is not suitable for data with non-linear patterns or non-spherical distributions.
4. K-Means is not suitable for data with outliers or noise.

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

Ans:

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: The cluster centroids for the second iteration would be:

C1: (4, 4);

C2: (2, 2);

C3: (7, 7).

The SSE for this clustering would be:

SSE = (2 − 4)2 + (2 − 4)2 + (6 − 4)2 + (6 − 4)2 + (0 − 2)2 + (4 − 2)2 + (0 − 2)2 + (4 − 2)2 + (0 − 2)2 + (4 − 2)2 + (0 − 2)2 + (4 − 2)2 + (5 − 7)2 + (9 − 7)2 = 32.

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 ?

Ans: Step 1: **Understanding the problem**

The problem is asking us to determine if software flaws discovered during testing are identical. We have 20 defect data points that are clustered into 5 clusters using the k-means algorithm. Any new defect formed after the 5 clusters of defects have been identified must be listed as one of the forms identified by clustering. We need to explain this process using a simple diagram.

Step 2: **Understanding the k-means algorithm**

Before we proceed with the problem, let's understand the k-means algorithm. The k-means algorithm is an unsupervised machine learning algorithm used for clustering. It groups data points into k clusters based on their similarity. The algorithm starts by selecting k random centroids and assigns each data point to its nearest centroid. It then calculates the new centroids based on the mean of the data points in each cluster and repeats the process until the centroids no longer change.

Step 3: **Applying the k-means algorithm**

In our problem, we have 20 defect data points that are clustered into 5 clusters using the k-means algorithm. We can represent this as follows:

Cluster 1: Defect data points 1, 4, 6, 9, 11

`Cluster 2:` Defect data points 2, 7, 8, 14

`Cluster 3:` Defect data points 3, 10, 12, 15, 18

`Cluster 4:` Defect data points 5, 13, 16, 17, 19

`Cluster 5:` Defect data point 20

we can see that the 20 defect data points have been grouped into 5 clusters based on their similarity using the k-means algorithm. Cluster 1 contains defect data points 1, 4, 6, 9, and 11, which are similar to each other. Cluster 2 contains defect data points 2, 7, 8, and 14, which are similar to each other. Cluster 3 contains defect data points 3, 10, 12, 15, and 18, which are similar to each other. Cluster 4 contains defect data points 5, 13, 16, 17, and 19, which are similar to each other. Finally, Cluster 5 contains only one defect data point, which means it is unique and not similar to any other defect.

`Step 4:` **Listing new defects**

As per the problem statement, 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 if a new defect is discovered during testing, it should be assigned to one of the existing clusters based on its similarity to the defect data points in that cluster.

For example, if a new defect is discovered during testing and it is similar to defect data points 1, 4, 6, 9, and 11, it should be assigned to Cluster 1. If it is similar to defect data points 2, 7, 8, and 14, it should be assigned to Cluster 2, and so on.

`Step 5:` **Conclusion**

In conclusion, we have solved the problem of determining if software flaws discovered during testing are identical. We used the k-means algorithm to cluster the defect data points into 5 clusters based on their similarity. We represented the clusters in a diagram and explained how to list any new defects that are discovered during testing.

In [ ]:

```
1
```