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.

```
a) Using the k-means method, we can create two clusters for each set of
centroids:

Set 1 Centroids: 15, 32
Initial Cluster Assignment:

Data points 5, 10, 15 are closer to centroid 15.
Data points 20, 25, 30, 35 are closer to centroid 32.
Updated Cluster Assignment:
Cluster 1: 5, 10, 15
Cluster 2: 20, 25, 30, 35

Set 2 Centroids: 12, 30
Initial Cluster Assignment:

Data points 5, 10, 15 are closer to centroid 12.
Data points 20, 25, 30, 35 are closer to centroid 30.
Updated Cluster Assignment:
Cluster 1: 5, 10, 15
Cluster 2: 20, 25, 30, 35

b) To calculate the Sum of Squared Errors (SSE) for each set of centroid
values, we need to compute the squared distance between
each data point and its assigned centroid, and then sum these squared
distances for all data points in each cluster.

Set 1 SSE calculation:
SSE1 = (5-15)^2 + (10-15)^2 + (15-15)^2 + (20-32)^2 + (25-32)^2 + (30-32)^2
+ (35-32)^2
= 100 + 25 + 0 + 144 + 49 + 4 + 9
= 331

Set 2 SSE calculation:
SSE2 = (5-12)^2 + (10-12)^2 + (15-12)^2 + (20-30)^2 + (25-30)^2 + (30-30)^2
+ (35-30)^2
= 49 + 4 + 9 + 100 + 25 + 0 + 25
= 212

Therefore, the SSE for Set 1 is 331 and the SSE for Set 2 is 212.
```

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

concepts ?

```
Market Basket Analysis uses association analysis concepts to uncover
patterns and relationships in customer purchasing behavior.
It identifies frequent itemsets and generates association rules based on the
co-occurrence of items in transactions. These rules reveal
the associations between products and help businesses understand customer
preferences, optimize product placement, and develop
targeted marketing strategies
```

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

In [ ]:

```
Apriori algorithm example:

Given a transaction dataset:

TID Items
1 Bread, Milk
2 Bread, Diapers, Beer
3 Milk, Diapers, Eggs
4 Bread, Milk, Diapers, Beer
5 Bread, Milk, Diapers, Eggs

Find frequent 1-itemsets:
{Bread}, {Milk}, {Diapers}

Generate candidate 2-itemsets:
{Bread, Milk}, {Bread, Diapers}, {Milk, Diapers}

Find frequent 2-itemsets:
{Bread, Milk}, {Bread, Diapers}, {Milk, Diapers}

Generate candidate 3-itemsets:
{Bread, Milk, Diapers}

Find frequent 3-itemsets:
{Bread, Milk, Diapers}

Generate association rules:
{Bread, Milk} => {Diapers}

These association rules indicate that if a customer buys bread and milk, they
```

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 typically
measured using a distance metric such as Euclidean distance,
Manhattan distance, or correlation distance. The choice of distance metric
depends on the nature of the data and the specific
requirements of the clustering task.

The iteration in hierarchical clustering continues until all data points are
assigned to a single cluster or until a stopping criterion
```

is met. One common stopping criterion is based on a threshold distance or a predetermined number of desired clusters.
The distance metric is used to decide when to end the iteration by comparing the distances between clusters at each step.

At each iteration, the algorithm merges the two closest clusters based on the chosen distance metric. The distance between clusters can be computed using different methods such as single linkage, complete linkage, or average linkage. The specific linkage method determines how the distance between clusters is calculated.

The iteration ends when the distance between the remaining clusters exceeds the defined threshold or when the desired number of clusters is reached. This process forms a hierarchical structure or a dendrogram, where the vertical axis represents the distance between clusters. By setting a threshold or desired number of clusters, the algorithm determines the appropriate stopping point and partitions the data accordingl

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

In the k-means algorithm, the recomputation of cluster centroids occurs in the following steps:

Initialization: Initially, k centroids are randomly chosen as the center points of the clusters.

Assignment: Each data point is assigned to the nearest centroid based on a chosen distance metric (usually Euclidean distance).

Update: After the assignment step, the cluster memberships are fixed. The centroids are updated by computing the mean of the data points within each cluster.

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

The elbow method is a popular technique for determining the required number of clusters in a clustering exercise.
It involves calculating the within-cluster sum of squares (WCSS) for different values of k and selecting the point on the plot where the decrease in WCSS levels off significantly as the optimal number of clusters.

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

In [ ]:

```
Advantages of the k-means algorithm:

Simple and easy to understand.
Computationally efficient and can handle large datasets.
Works well with spherical and well-separated clusters.
Converges to a local optimum, which can be sufficient for many practical appli

Disadvantages of the k-means algorithm:

Requires the number of clusters (k) to be specified in advance.
Sensitive to initial centroid selection, which can lead to different results.
Assumes clusters of similar size and density.
Not suitable for non-linear or irregularly shaped clusters.
Can be affected by outliers, as they can disproportionately influence cluster
Overall, while the k-means algorithm has its limitations, it remains a popular
simplicity and efficiency.
```

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

```
A clustering diagram consists of scattered data points in a two-dimensional
space, with each point representing a data sample.
Clusters are formed by grouping together data points that are close to each
other, and each cluster is represented by a distinct
shape or boundary. The goal is to identify natural groupings or patterns in
the data based on proximity
```

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?

```
In the given scenario, the team is using the k-means algorithm to cluster 20
defect data points into 5 clusters.
The k-means algorithm starts by randomly initializing 5 cluster centroids.
It then iteratively assigns each data point to
the nearest centroid and updates the centroids based on the assigned points.
This process continues until convergence, where the
centroids remain unchanged.

Once the clustering process is complete, each defect data point will be
assigned to one of the 5 identified clusters.
Any new defect that arises after clustering must also be assigned to one of
these clusters based on its similarity to the
existing clusters. This ensures that all defects are categorized into one of
the predefined forms identified by clustering.
```

```
The diagram would visually represent the 20 defect data points scattered in
a space and the resulting clusters formed by the k-means
algorithm. Each data point would be connected to the centroid of its
assigned cluster. However, as a text-based interface,
I'm unable to provide a diagram here.
```

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 ?

In [1]:

```
import numpy as np
from sklearn.cluster import KMeans

# Defect data points
defects = np.array([[x, x] for x in range(20)])

# K-means clustering with 5 clusters
kmeans = KMeans(n_clusters=5, random_state=0)
kmeans.fit(defects)

# Cluster assignments for each data point
cluster_labels = kmeans.labels_

# Centroids of the clusters
centroids = kmeans.cluster_centers_

# New defect
new_defect = np.array([[25, 25]])

# Predicting the cluster for the new defect
new_defect_cluster = kmeans.predict(new_defect)

# Print cluster assignments and centroids
print("Cluster Assignments:")
print(cluster_labels)
print("\nCentroids:")
print(centroids)
print("\nNew Defect Cluster:")
print(new_defect_cluster)
```

```
Cluster Assignments:
[3 3 3 1 1 1 1 4 4 4 4 0 0 0 0 0 2 2 2 2]

Centroids:
[[13.  13. ]
 [ 4.5  4.5]
 [17.5 17.5]
 [ 1.   1. ]
 [ 8.5  8.5]]

New Defect Cluster:
[2]
```

In [ ]: