Explain the purpose of “Attribute selection measures” in classification by decision tree induction? How we can use the “Tree pruning” in classification?

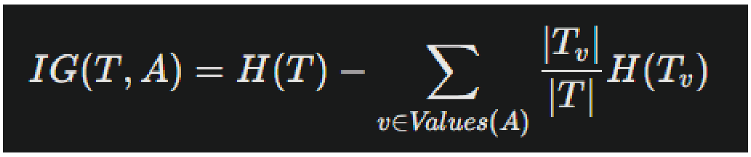
In classification by decision tree induction, **attribute selection measures** and **tree pruning** are essential techniques to improve the accuracy and simplicity of the model.

Attribute Selection Measures

The purpose of attribute selection measures (also known as *splitting criteria*) in decision tree induction is to identify which attribute (or feature) should be used at each node to split the dataset. A good attribute for splitting helps increase the purity of the resulting subsets, which ideally should contain mostly instances from the same class.

Common attribute selection measures include:

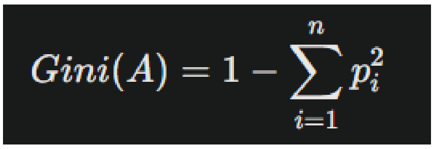
1. **Information Gain**: Based on entropy, information gain measures how much information an attribute gives about the class. Attributes with high information gain are chosen as they reduce the impurity of the splits.



$$IG(T, A) = H(T) - \sum_{v \in Values(A)} \frac{|T_v|}{|T|} H(T_v)$$

where $H(T)$ is the entropy of the target variable T, and T_v represents the subset of T for which attribute A has value v

2. **Gini Index:** This measure assesses impurity in a dataset. It ranges from 0 (perfectly pure) to 1 (maximum impurity). The Gini index for an attribute is calculated as:



$$Gini(A) = 1 - \sum_{i=1}^{n} p_i^2$$

where pi is the proportion of class i in the dataset.

3. **Gain Ratio**: A variant of information gain that accounts for biases by normalizing against the intrinsic information of the split. It’s used to avoid the selection of attributes with many values that can lead to overfitting.

4. **Chi-Square:** This measure tests the independence of an attribute with respect to the class label. A high chi-square value indicates that the attribute is highly dependent on the class.

By selecting attributes based on these measures, the decision tree is able to split data more effectively, resulting in a model that is more accurate and interpretable.

Tree Pruning

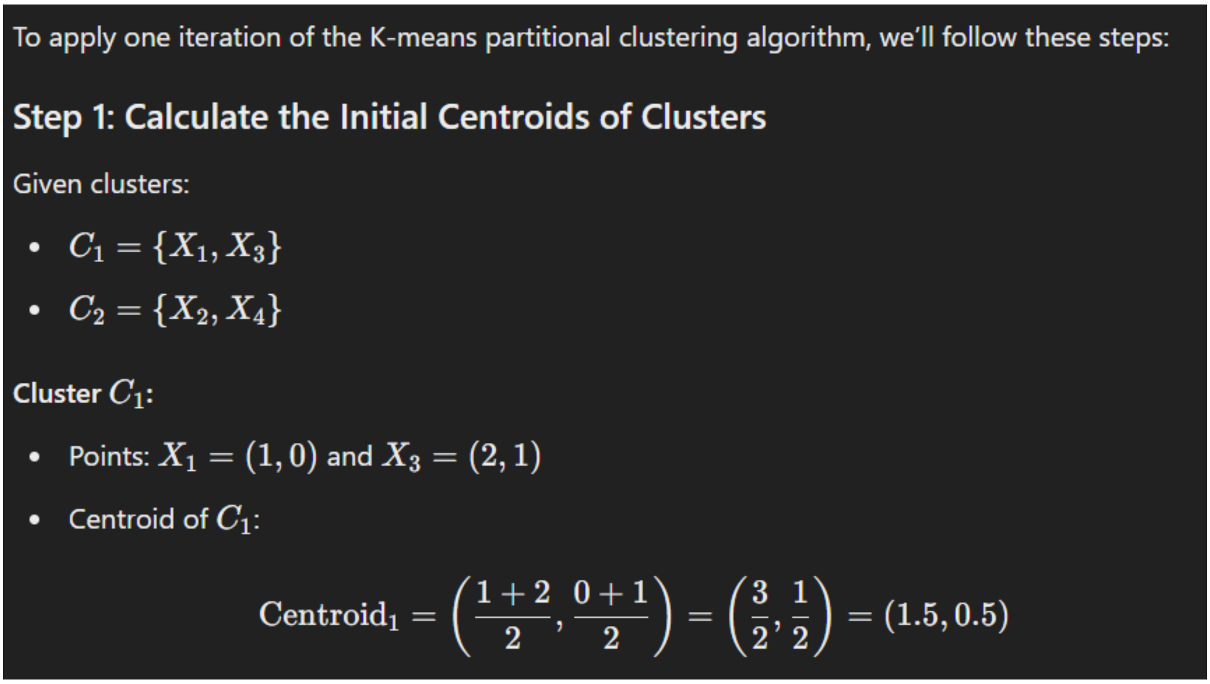
Tree pruning is a technique to reduce the complexity of a decision tree, enhancing its generalizability and minimizing the risk of overfitting. During training, a decision tree may grow excessively complex, capturing noise and irrelevant details that can decrease its performance on new data.

Pruning can be applied in two main ways:

1. **Pre-pruning (Early Stopping):** This involves setting conditions to stop the tree from growing too large, such as limiting the depth of the tree, specifying a minimum number of samples required to split a node, or setting a threshold for minimum information gain. These conditions prevent overfitting by restricting tree growth.
2. **Post-pruning:** After building the full tree, post-pruning removes branches or nodes that do not contribute significantly to classification accuracy. Methods include:
    - **Cost Complexity Pruning:** This removes nodes or subtrees based on a cost function that balances tree complexity and classification accuracy.
    - **Reduced Error Pruning:** It removes nodes if the removal does not increase error on a validation dataset.

Pruning simplifies the tree, making it easier to interpret while often improving its accuracy on unseen data by focusing on the most relevant patterns in the dataset.

Given the samples X1 = {1, 0}, X2 = {0, 1}, X3 = {2, 1}, and X4 = {3, 3}, suppose that the samples are randomly clustered into two clusters C1 = {X1, X3} and C2 = {X2, X4}. Apply one iteration of the K-means partitional clustering algorithm, and find a new distribution of samples in clusters.



To apply one iteration of the K-means partitional clustering algorithm, we'll follow these steps:

Step 1: Calculate the Initial Centroids of Clusters

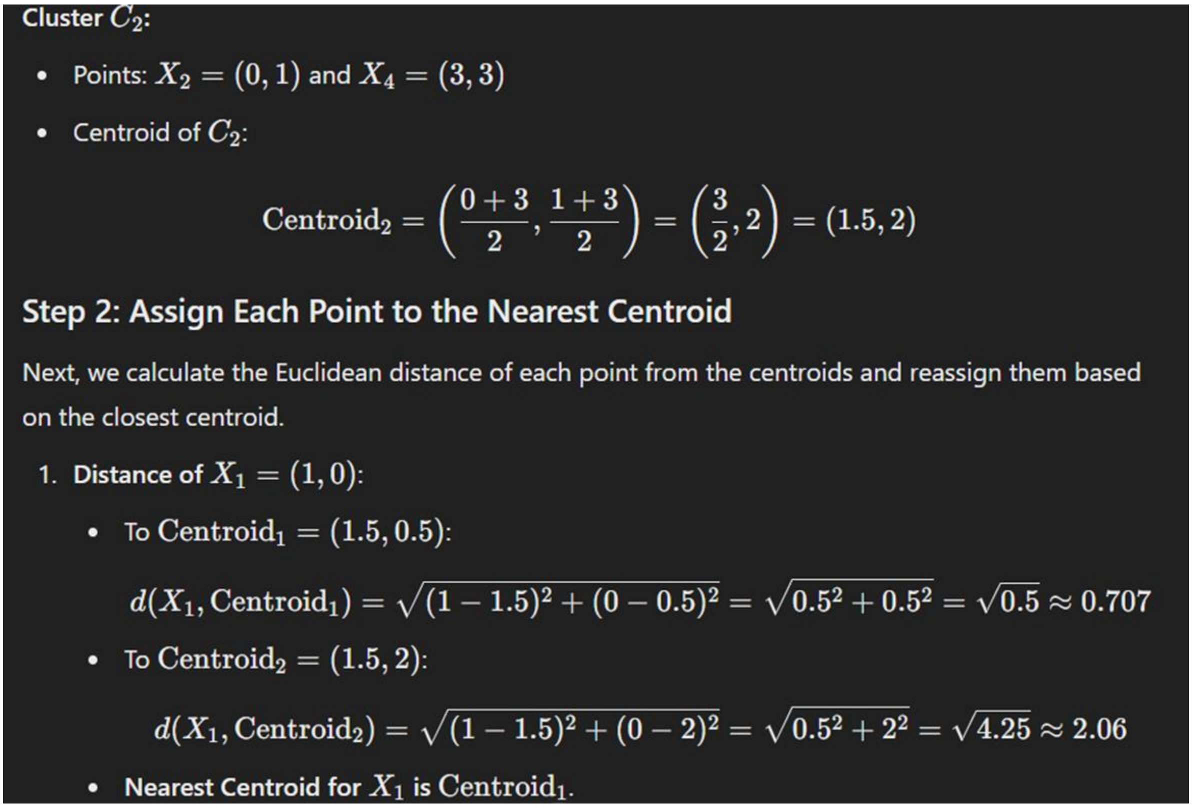
Given clusters:

- $C_1 = \{X_1, X_3\}$
- $C_2 = \{X_2, X_4\}$

Cluster C_1:

- Points: $X_1 = (1, 0)$ and $X_3 = (2, 1)$
- Centroid of C_1:

$$\text{Centroid}_1 = \left(\frac{1+2}{2}, \frac{0+1}{2}\right) = \left(\frac{3}{2}, \frac{1}{2}\right) = (1.5, 0.5)$$



Cluster C_2:

- Points: $X_2 = (0, 1)$ and $X_4 = (3, 3)$
- Centroid of C_2:

$$\text{Centroid}_2 = \left(\frac{0+3}{2}, \frac{1+3}{2}\right) = \left(\frac{3}{2}, 2\right) = (1.5, 2)$$

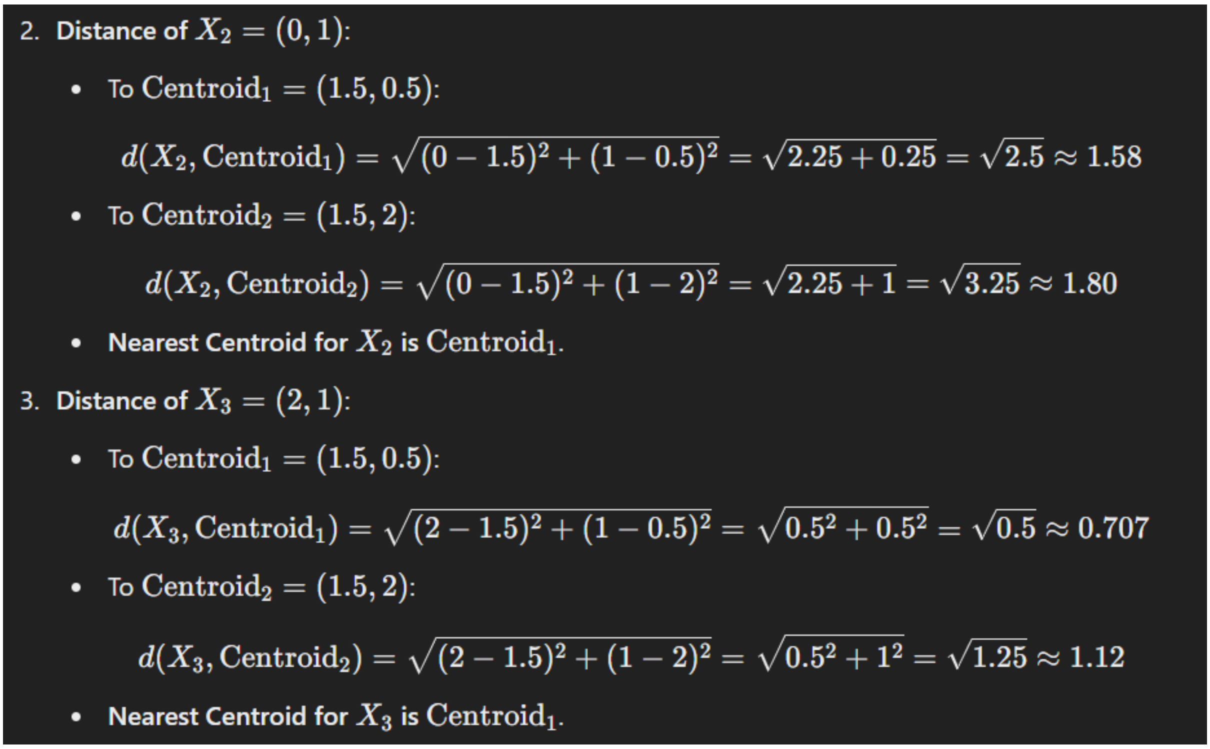
Step 2: Assign Each Point to the Nearest Centroid

Next, we calculate the Euclidean distance of each point from the centroids and reassign them based on the closest centroid.

1. **Distance of $X_1 = (1, 0)$:**
   - To $\text{Centroid}_1 = (1.5, 0.5)$:

     $$d(X_1, \text{Centroid}_1) = \sqrt{(1-1.5)^2 + (0-0.5)^2} = \sqrt{0.5^2 + 0.5^2} = \sqrt{0.5} \approx 0.707$$
   - To $\text{Centroid}_2 = (1.5, 2)$:

     $$d(X_1, \text{Centroid}_2) = \sqrt{(1-1.5)^2 + (0-2)^2} = \sqrt{0.5^2 + 2^2} = \sqrt{4.25} \approx 2.06$$
   - **Nearest Centroid for X_1 is Centroid_1.**



2. **Distance of $X_2 = (0, 1)$:**
   - To $\text{Centroid}_1 = (1.5, 0.5)$:

     $$d(X_2, \text{Centroid}_1) = \sqrt{(0 - 1.5)^2 + (1 - 0.5)^2} = \sqrt{2.25 + 0.25} = \sqrt{2.5} \approx 1.58$$

   - To $\text{Centroid}_2 = (1.5, 2)$:

     $$d(X_2, \text{Centroid}_2) = \sqrt{(0 - 1.5)^2 + (1 - 2)^2} = \sqrt{2.25 + 1} = \sqrt{3.25} \approx 1.80$$

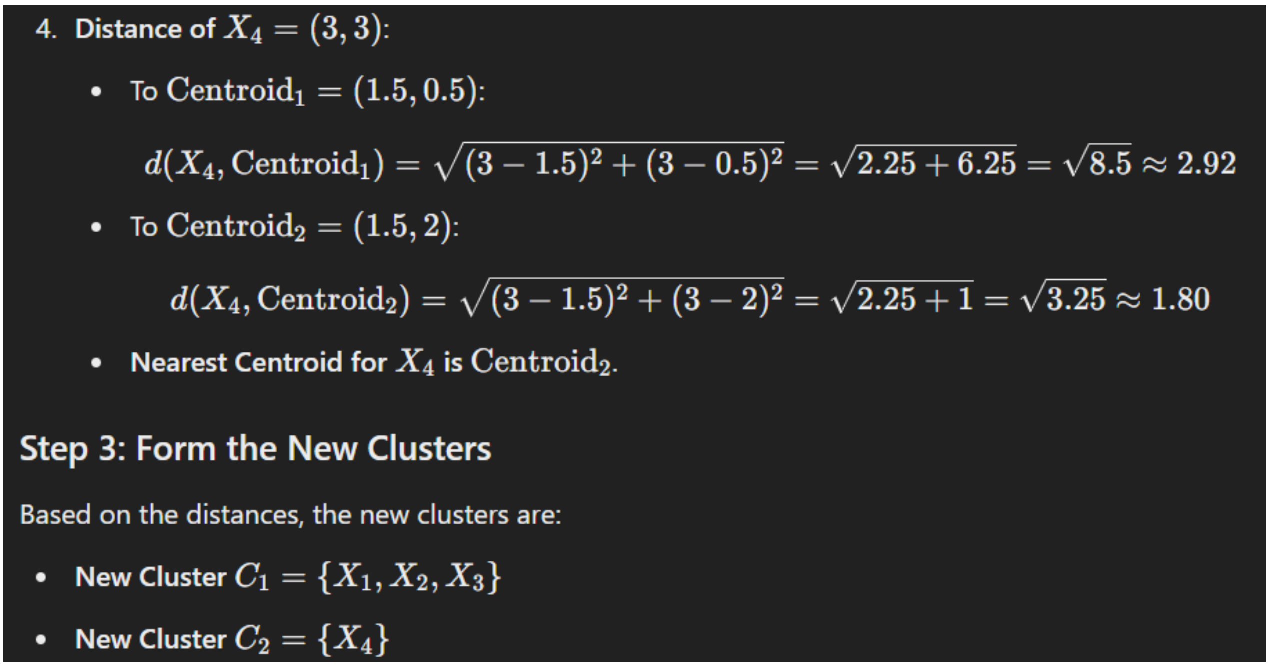
   - **Nearest Centroid for X_2 is Centroid_1.**
3. **Distance of $X_3 = (2, 1)$:**
   - To $\text{Centroid}_1 = (1.5, 0.5)$:

     $$d(X_3, \text{Centroid}_1) = \sqrt{(2 - 1.5)^2 + (1 - 0.5)^2} = \sqrt{0.5^2 + 0.5^2} = \sqrt{0.5} \approx 0.707$$

   - To $\text{Centroid}_2 = (1.5, 2)$:

     $$d(X_3, \text{Centroid}_2) = \sqrt{(2 - 1.5)^2 + (1 - 2)^2} = \sqrt{0.5^2 + 1^2} = \sqrt{1.25} \approx 1.12$$

   - **Nearest Centroid for X_3 is Centroid_1.**



4. **Distance of $X_4 = (3, 3)$:**
   - To $\text{Centroid}_1 = (1.5, 0.5)$:

     $$d(X_4, \text{Centroid}_1) = \sqrt{(3 - 1.5)^2 + (3 - 0.5)^2} = \sqrt{2.25 + 6.25} = \sqrt{8.5} \approx 2.92$$

   - To $\text{Centroid}_2 = (1.5, 2)$:

     $$d(X_4, \text{Centroid}_2) = \sqrt{(3 - 1.5)^2 + (3 - 2)^2} = \sqrt{2.25 + 1} = \sqrt{3.25} \approx 1.80$$

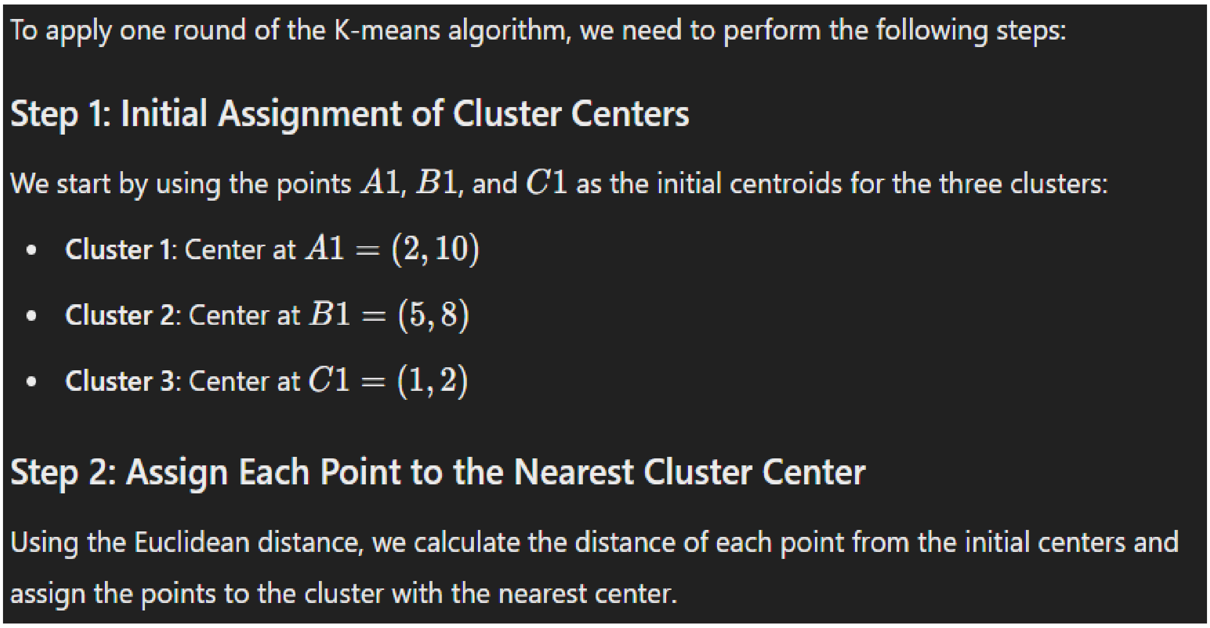
   - **Nearest Centroid for X_4 is Centroid_2.**

Step 3: Form the New Clusters

Based on the distances, the new clusters are:

- **New Cluster $C_1 = \{X_1, X_2, X_3\}$**
- **New Cluster $C_2 = \{X_4\}$**

Suppose that the data-mining task is to cluster the following eight points (representing location) into three clusters: A1 (2,10) ; A2 (2,5) ; A3 (8,4) ; B1 (5,8) ; B2 (7,5) ; B3 (6,4) ; C1 (1,2) ; C2 (4,9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to determine: the three cluster centers after the first round of execution.



To apply one round of the K-means algorithm, we need to perform the following steps:

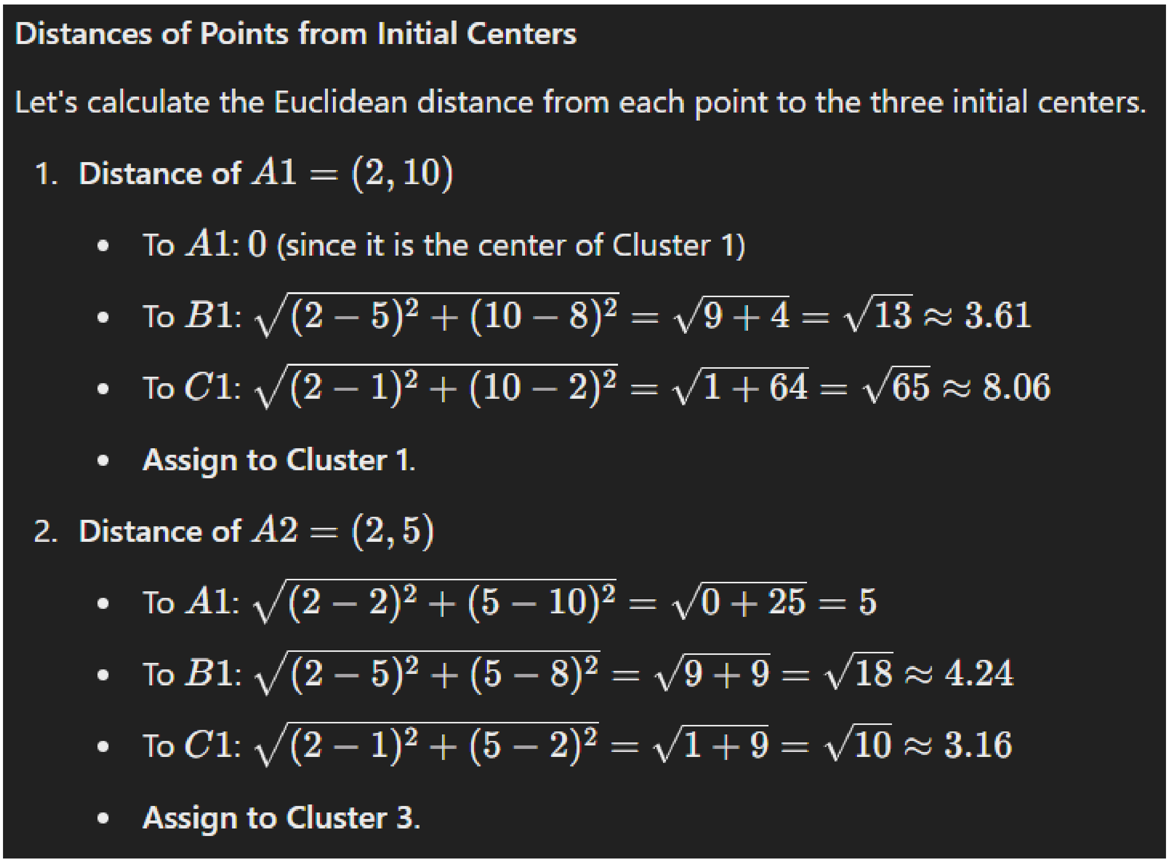
Step 1: Initial Assignment of Cluster Centers

We start by using the points $A1$, $B1$, and $C1$ as the initial centroids for the three clusters:

- **Cluster 1**: Center at $A1 = (2, 10)$
- **Cluster 2**: Center at $B1 = (5, 8)$
- **Cluster 3**: Center at $C1 = (1, 2)$

Step 2: Assign Each Point to the Nearest Cluster Center

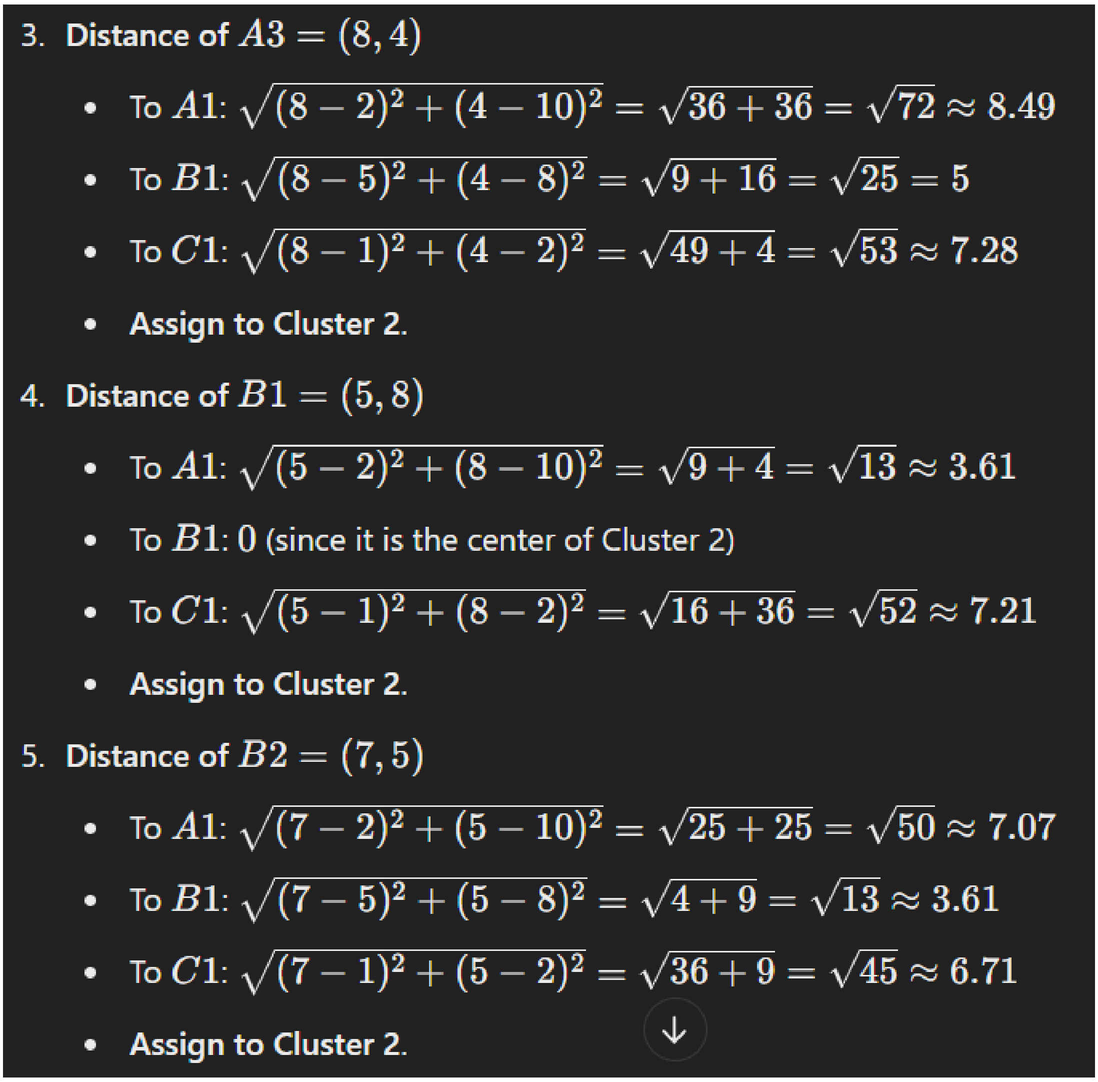
Using the Euclidean distance, we calculate the distance of each point from the initial centers and assign the points to the cluster with the nearest center.



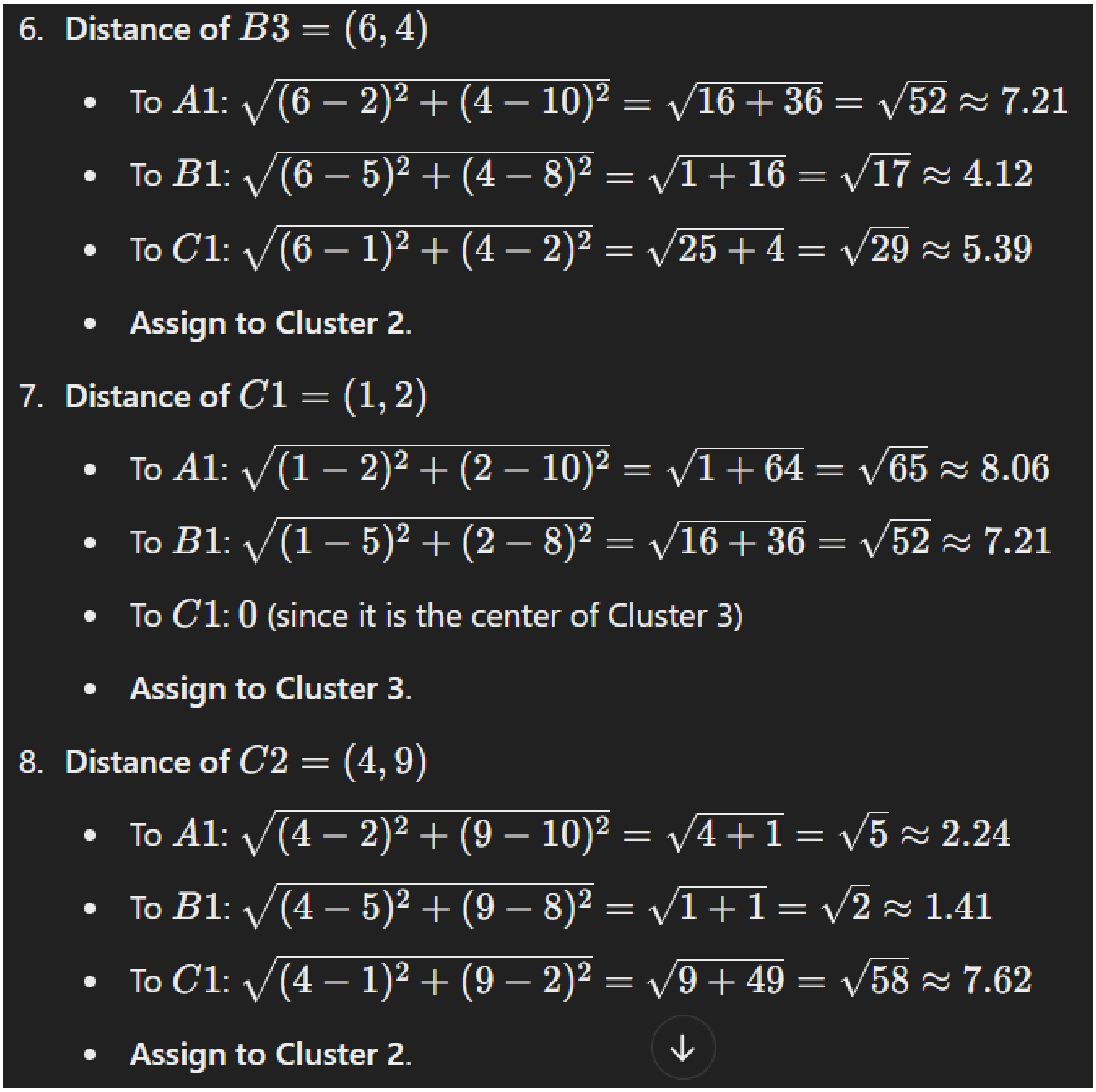
Distances of Points from Initial Centers

Let's calculate the Euclidean distance from each point to the three initial centers.

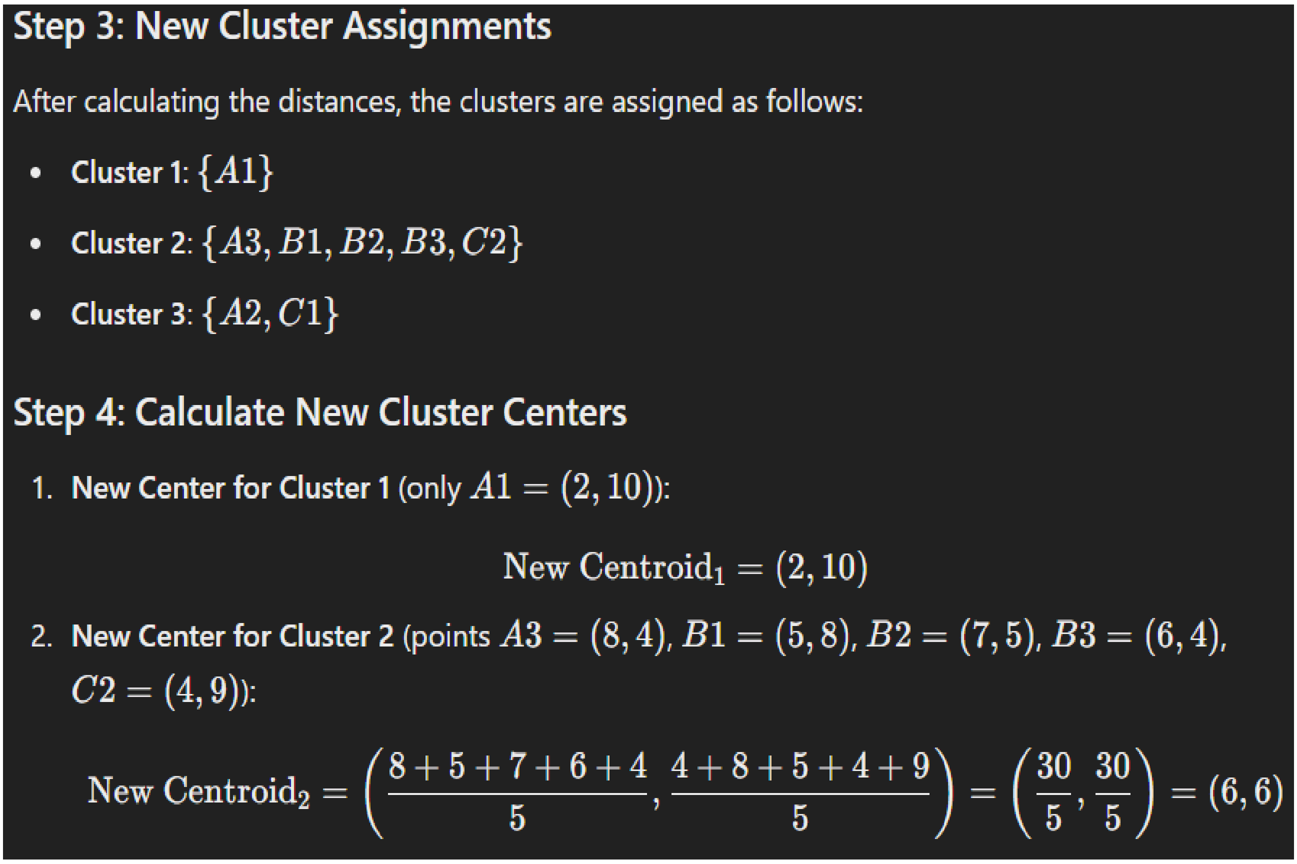
1. **Distance of $A1 = (2, 10)$**
   - To $A1$: 0 (since it is the center of Cluster 1)
   - To $B1$: $\sqrt{(2-5)^2 + (10-8)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61$
   - To $C1$: $\sqrt{(2-1)^2 + (10-2)^2} = \sqrt{1+64} = \sqrt{65} \approx 8.06$
   - **Assign to Cluster 1.**
2. **Distance of $A2 = (2, 5)$**
   - To $A1$: $\sqrt{(2-2)^2 + (5-10)^2} = \sqrt{0+25} = 5$
   - To $B1$: $\sqrt{(2-5)^2 + (5-8)^2} = \sqrt{9+9} = \sqrt{18} \approx 4.24$
   - To $C1$: $\sqrt{(2-1)^2 + (5-2)^2} = \sqrt{1+9} = \sqrt{10} \approx 3.16$
   - **Assign to Cluster 3.**



3. Distance of $A3 = (8, 4)$
   - To $A1$: $\sqrt{(8-2)^2 + (4-10)^2} = \sqrt{36+36} = \sqrt{72} \approx 8.49$
   - To $B1$: $\sqrt{(8-5)^2 + (4-8)^2} = \sqrt{9+16} = \sqrt{25} = 5$
   - To $C1$: $\sqrt{(8-1)^2 + (4-2)^2} = \sqrt{49+4} = \sqrt{53} \approx 7.28$
   - **Assign to Cluster 2.**
4. Distance of $B1 = (5, 8)$
   - To $A1$: $\sqrt{(5-2)^2 + (8-10)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61$
   - To $B1$: 0 (since it is the center of Cluster 2)
   - To $C1$: $\sqrt{(5-1)^2 + (8-2)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$
   - **Assign to Cluster 2.**
5. Distance of $B2 = (7, 5)$
   - To $A1$: $\sqrt{(7-2)^2 + (5-10)^2} = \sqrt{25+25} = \sqrt{50} \approx 7.07$
   - To $B1$: $\sqrt{(7-5)^2 + (5-8)^2} = \sqrt{4+9} = \sqrt{13} \approx 3.61$
   - To $C1$: $\sqrt{(7-1)^2 + (5-2)^2} = \sqrt{36+9} = \sqrt{45} \approx 6.71$
   - **Assign to Cluster 2.**



6. Distance of $B3 = (6, 4)$
   - To $A1$: $\sqrt{(6-2)^2 + (4-10)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$
   - To $B1$: $\sqrt{(6-5)^2 + (4-8)^2} = \sqrt{1+16} = \sqrt{17} \approx 4.12$
   - To $C1$: $\sqrt{(6-1)^2 + (4-2)^2} = \sqrt{25+4} = \sqrt{29} \approx 5.39$
   - **Assign to Cluster 2.**
7. Distance of $C1 = (1, 2)$
   - To $A1$: $\sqrt{(1-2)^2 + (2-10)^2} = \sqrt{1+64} = \sqrt{65} \approx 8.06$
   - To $B1$: $\sqrt{(1-5)^2 + (2-8)^2} = \sqrt{16+36} = \sqrt{52} \approx 7.21$
   - To $C1$: 0 (since it is the center of Cluster 3)
   - **Assign to Cluster 3.**
8. Distance of $C2 = (4, 9)$
   - To $A1$: $\sqrt{(4-2)^2 + (9-10)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.24$
   - To $B1$: $\sqrt{(4-5)^2 + (9-8)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.41$
   - To $C1$: $\sqrt{(4-1)^2 + (9-2)^2} = \sqrt{9+49} = \sqrt{58} \approx 7.62$
   - **Assign to Cluster 2.**



Step 3: New Cluster Assignments

After calculating the distances, the clusters are assigned as follows:

- **Cluster 1**: $\{A1\}$
- **Cluster 2**: $\{A3, B1, B2, B3, C2\}$
- **Cluster 3**: $\{A2, C1\}$

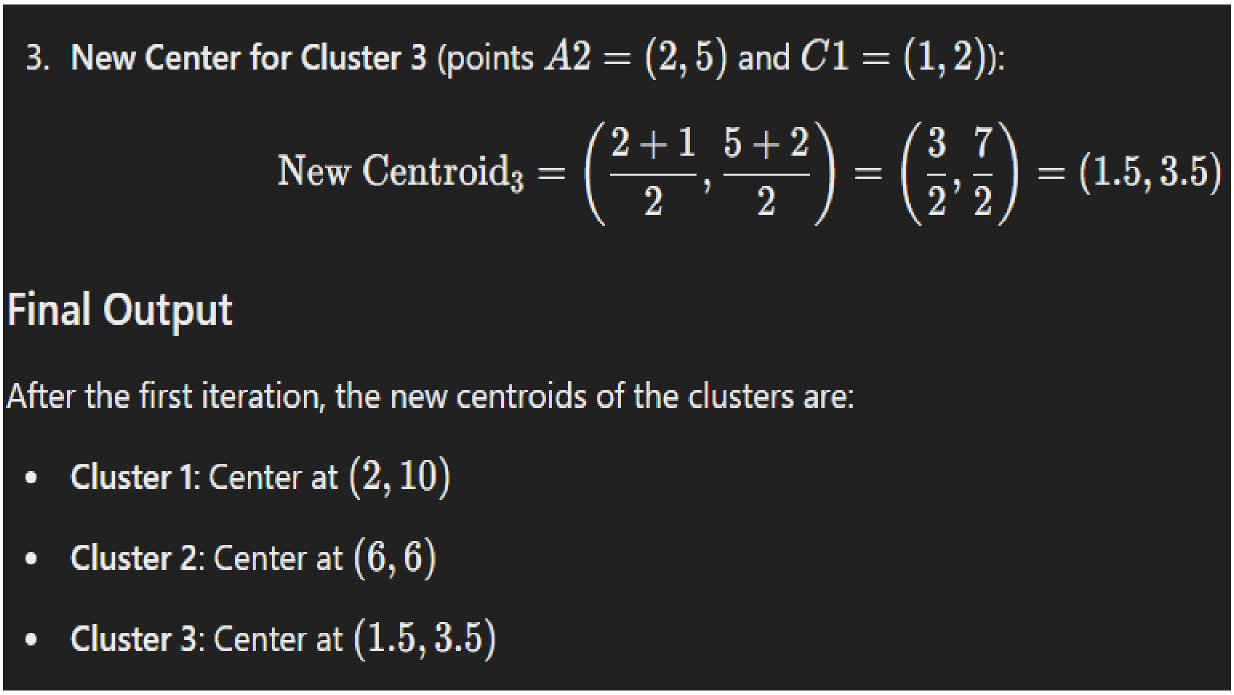
Step 4: Calculate New Cluster Centers

1. **New Center for Cluster 1** (only $A1 = (2, 10)$):

$$\text{New Centroid}_1 = (2, 10)$$

2. **New Center for Cluster 2** (points $A3 = (8, 4)$, $B1 = (5, 8)$, $B2 = (7, 5)$, $B3 = (6, 4)$, $C2 = (4, 9)$):

$$\text{New Centroid}_2 = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = \left(\frac{30}{5}, \frac{30}{5}\right) = (6, 6)$$



3. **New Center for Cluster 3** (points $A2 = (2, 5)$ and $C1 = (1, 2)$):

$$\text{New Centroid}_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = \left(\frac{3}{2}, \frac{7}{2}\right) = (1.5, 3.5)$$

Final Output

After the first iteration, the new centroids of the clusters are:

- **Cluster 1**: Center at $(2, 10)$
- **Cluster 2**: Center at $(6, 6)$
- **Cluster 3**: Center at $(1.5, 3.5)$

Compare and contrast the advantages and disadvantages of K-means and bisecting K-means in terms of performance and quality of clustering results.

K-means and bisecting K-means are both popular clustering algorithms, each with its own advantages and disadvantages regarding performance and clustering quality. Here's a comparative analysis of the two:

K-means Clustering

Advantages:

1. **Simplicity**: K-means is straightforward to understand and implement, making it accessible for beginners in clustering.
2. **Efficiency**: It operates efficiently with a time complexity of O(n·k·i), where nis the number of data points, k is the number of clusters, and iii is the number of iterations.
3. **Scalability**: K-means can handle large datasets effectively, making it suitable for applications with a significant amount of data.
4. **Speed**: The algorithm converges relatively quickly, often requiring fewer iterations to reach a solution compared to more complex clustering methods.

Disadvantages:

1. **Sensitivity to Initialization**: The choice of initial centroids can significantly influence the results, leading to different clustering outcomes with different runs.
2. **Fixed Number of Clusters**: K-means requires prior knowledge of the number of clusters (k), which can be challenging when the optimal number is unknown.
3. **Assumes Spherical Clusters**: The algorithm tends to perform poorly on datasets with non-globular clusters or varying densities, as it assumes that clusters are spherical and evenly sized.
4. **Outlier Sensitivity**: K-means can be affected by outliers, which may distort the centroid positions and negatively impact clustering quality.

Bisecting K-means

Advantages:

1. **Improved Clustering Quality**: Bisecting K-means often produces higher-quality clusters, particularly in datasets with varying densities, by splitting clusters in a hierarchical manner.
2. **Better Initialization**: This method reduces sensitivity to initial placements, leading to more stable and reliable results.
3. **Hierarchical Structure**: The algorithm provides a clearer interpretation of the clustering results, allowing users to understand the relationships between clusters better.
4. **Flexibility**: Bisecting K-means adapts more effectively to varying shapes and sizes of clusters, making it suitable for diverse datasets.

Disadvantages:

1. **Higher Computational Cost**: The time complexity is O(n·k·i·logk), making it more computationally intensive compared to K-means, especially with large datasets.
2. **Complex Implementation**: The algorithm is more complicated to implement than K-means, which may require additional coding and understanding of hierarchical clustering concepts.
3. **Fixed Number of Final Clusters**: Like K-means, bisecting K-means also requires a predetermined k, which can still be a challenge when the optimal number of clusters is unknown.
4. **Potential for Inaccurate Splits**: The quality of clustering can be influenced by the initial choices made during the splitting process, leading to suboptimal results if not carefully executed.

Summary

- **Performance**: K-means is typically faster and more scalable, making it suitable for large datasets. In contrast, bisecting K-means may take longer to compute but often yields better quality clusters, particularly when dealing with complex shapes and varying densities.
- **Clustering Quality**: Bisecting K-means generally outperforms K-means in producing meaningful clusters, especially in datasets where the clusters do not conform to simple geometric shapes.