Unsupervised Learning

Outline

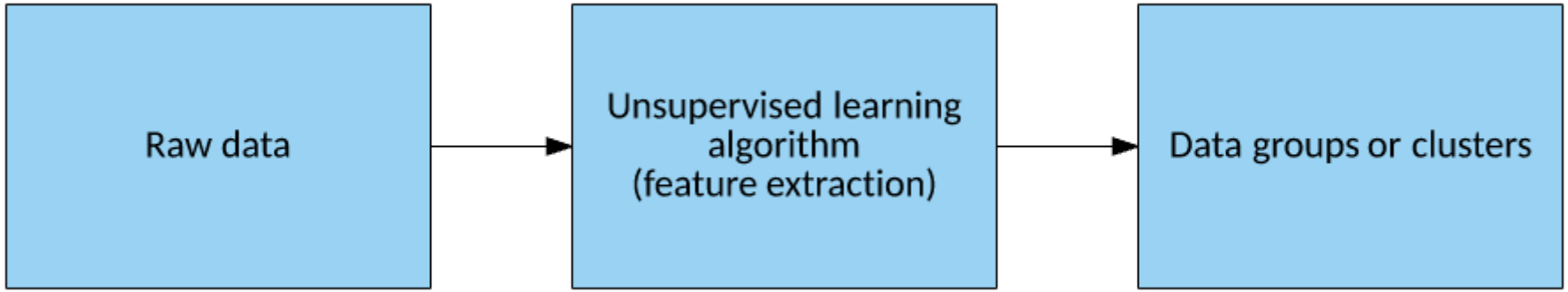
1. Introduction to Unsupervised learning
2. Types of Clustering and clustering algorithms
3. K-means algorithm
4. Hierarchical algorithm
5. Applications
6. Quiz

Unsupervised Learning

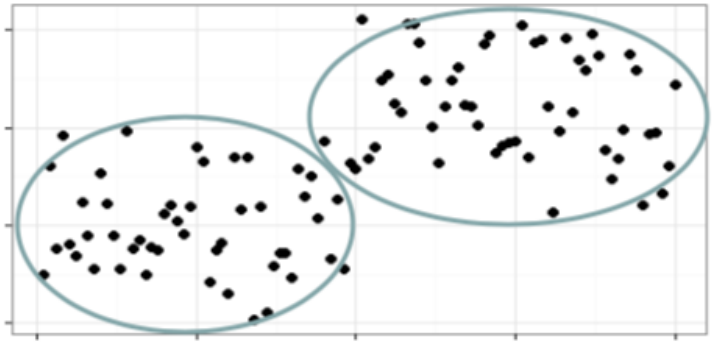
In **unsupervised learning**, an algorithm segregates the data in a data set in which the data is unlabeled based on some hidden features in the data.

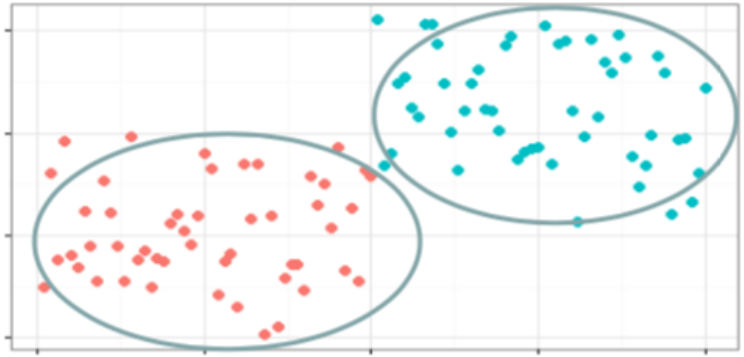
This function can be useful for discovering the hidden structure of data and for tasks like anomaly detection.

Unsupervised Learning



Unsupervised Learning





Clustering

Finding similarity groups in data are called **clusters**

- Data instances that are similar to (near) each other are in the same cluster.
- Data instances that are very different (far away) from each other fall in different clusters

Types of Clustering

Clustering can be divided into two subgroups :

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not.
- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.

Types of Clustering Algorithms

Clustering algorithms can be categorized based on their cluster model, that is based on how they form clusters or groups.

Types of Clustering Algorithms

Connectivity-based clustering: data points that are closer in the data space are more related (similar) than to data points farther away.

The clusters are formed by connecting data points according to their distance.

Examples - **hierarchical clustering** algorithm

Types of Clustering Algorithms

Centroid models: These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters.

k-means is a centroid based clustering

Types of Clustering Algorithms

Distribution-based clustering - Clustering is based on the notion of how probable is it for a data point to belong to a certain distribution, such as the Gaussian distribution.

Data points in a cluster belong to the same distribution. These models have a strong theoretical foundation, however they often suffer from overfitting.

Gaussian mixture models, using the expectation-maximization algorithm is a famous distribution based clustering method.

Types of Clustering Algorithms

Density-based methods search the data space for areas of varied density of data points. Clusters are defined as areas of higher density within the data space compared to other regions.

DBSCAN and **OPTICS** are some prominent density based clustering.

Quality of Clusters

Intra-cluster cohesion (compactness): Cohesion measures how near the data points in a cluster are to the cluster centroid.

Inter-cluster separation (isolation) - Separation means that different cluster centroids should be far away from each other.

Which Algorithm to use?

There is no ONE algorithm to rule them all !!

Clustering is an subjective task and there can be more than one correct clustering algorithm.

Every algorithm follows a different set of rules for defining the 'similarity' among data points.

K-Means Clustering

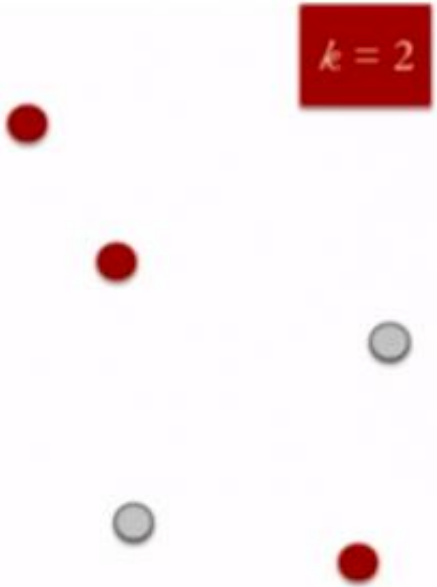
K-Means Clustering

K-means clustering is the most commonly used unsupervised machine learning algorithm for dividing a given dataset into k clusters.

K represents the number of clusters and must be provided by the user.

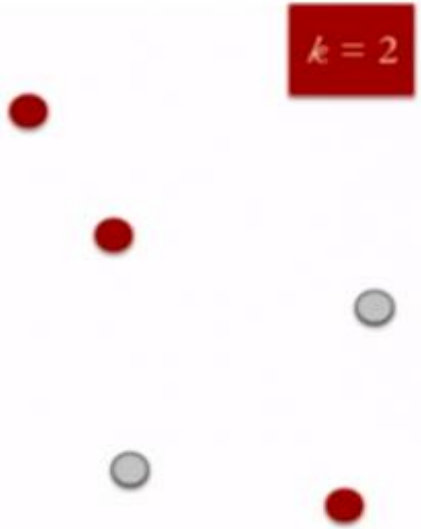
K-Means Clustering

Step 1 : Specify the desired number of clusters K. Let us choose k=2 for these 5 data points.



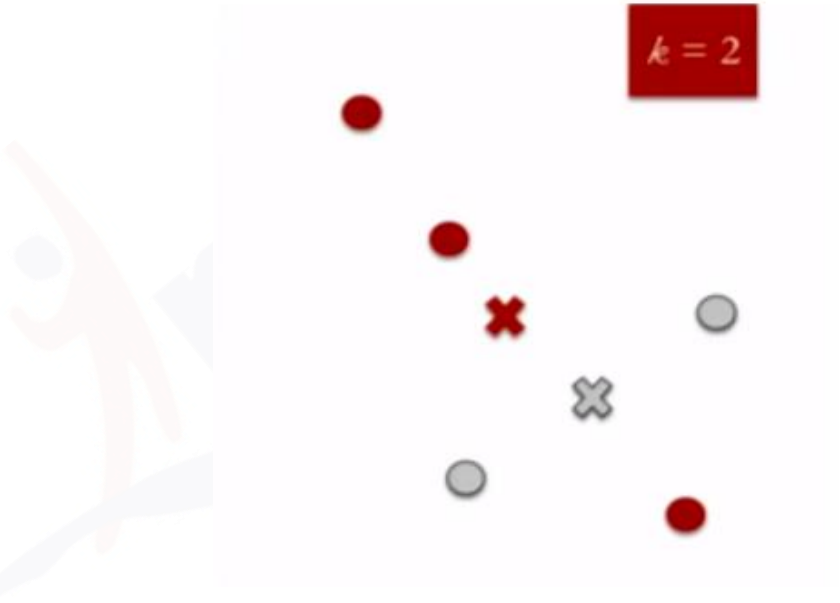
K-Means Clustering

Step 2 : Randomly assign each data point to a cluster : Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.



K-Means Clustering

Step 3 : Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.

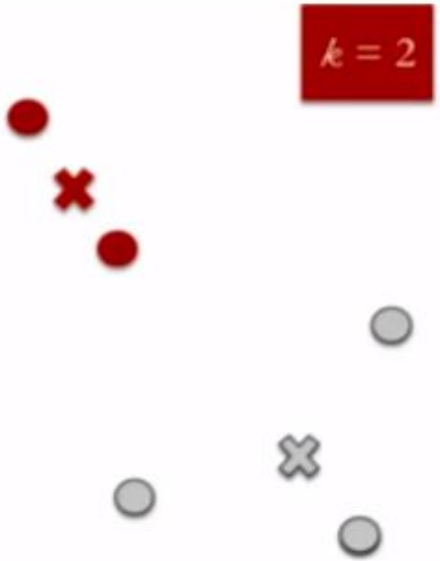


K-Means Clustering

Step 4 : Re-assign each point to the closest cluster centroid : Only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster

K-Means Clustering

Step 5 : Re-compute cluster centroids : Re-computing the centroids for both the clusters.



K-Means Clustering

Step 6 : Repeat steps 4 and 5 until no improvements are possible ~ When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm.

K-Means Clustering

K-Means runs on distance calculations, which uses "**Euclidean Distance**"

$$\text{Euclidean Distance} = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

K-Means Clustering

The basic restriction for K-Means algorithm is that your data should be continuous in nature.

It won’t work if data is categorical in nature!

K-Means Clustering - How many clusters(k) ?

Run k-means multiple times to see how model quality changes as the number of clusters changes.

Plots displaying this information help to determine the number of clusters and are often referred to as ***scree plots***.

K-Means Clustering - How many clusters(k) ?

The ideal plot will have an ***elbow*** where the quality measure improves more slowly as the number of clusters increases.

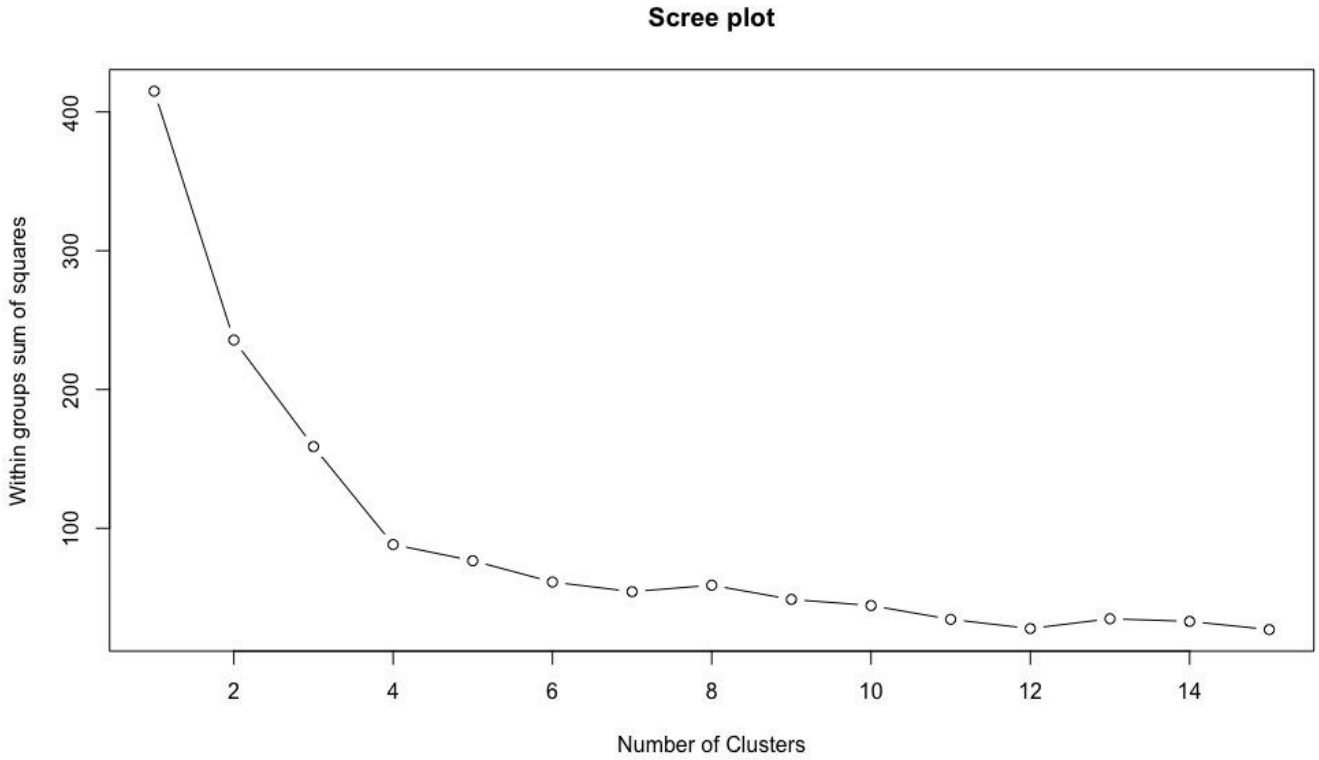
This indicates that the quality of the model is no longer improving substantially as the model complexity (i.e. number of clusters) increases.

In other words, the elbow indicates the number of clusters inherent in the data.

K-Means Clustering - How many clusters(k) ?

1. Compute k-means clustering for different values of k. For instance, by varying k from 1 to 15 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (**knee**) in the plot is generally considered as an indicator of the appropriate number of clusters.

K-Means Clustering - How many clusters(k) ?



K-Means Clustering in R

kmeans() function in R

Kmeans() output generates -

cluster: a vector of integers (from 1:k) indicating the cluster to which each point is allocated.

centers: a matrix of cluster centers.

K-Means Clustering in R

withinss: vector of within-cluster sum of squares, one component per cluster.

tot.withinss: total within-cluster sum of squares. That is, sum(withinss).

size: the number of points in each cluster.

Hierarchical Clustering

Hierarchical Clustering

Use distance matrix as clustering criteria.

This method does not need the number of clusters k as an input

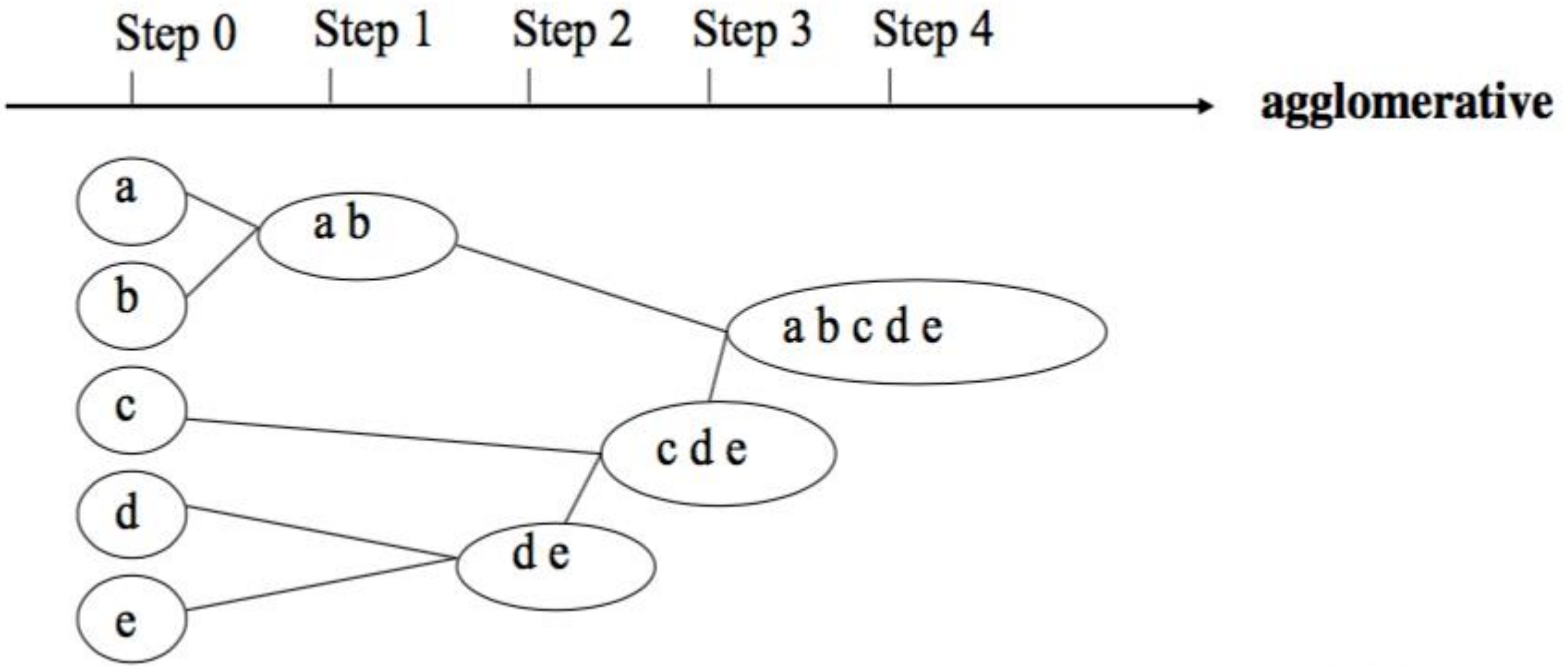
Hierarchical Clustering

This algorithm starts with all the data points assigned to a cluster of their own.

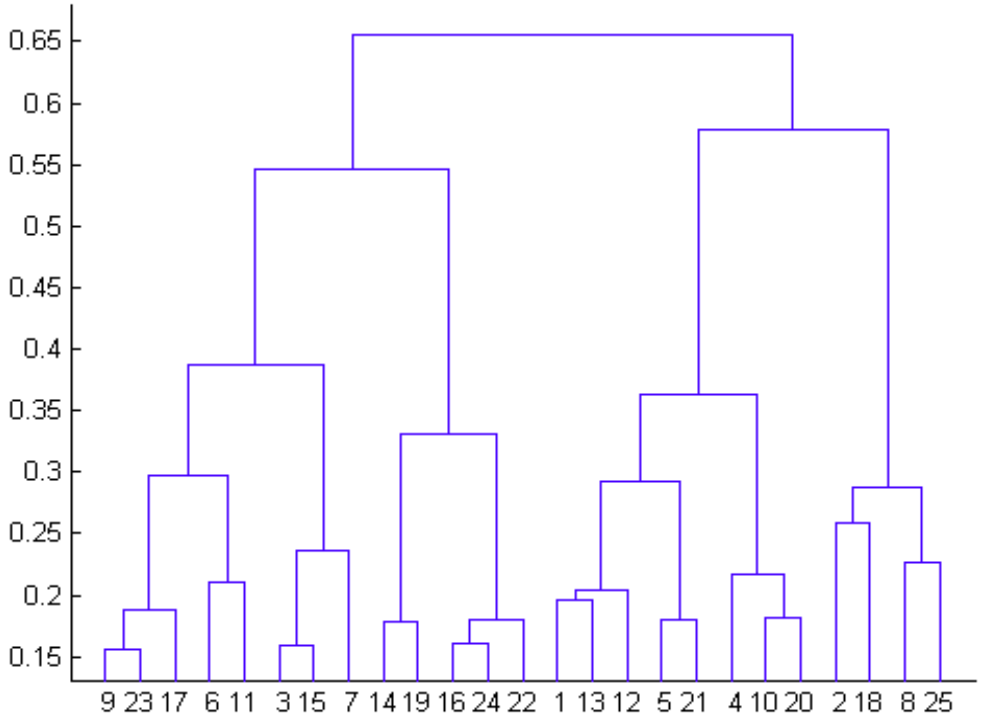
Then two nearest clusters are merged into the same cluster.

In the end, this algorithm terminates when there is only a single cluster left.

Hierarchical Clustering



Hierarchical Clustering - Dendrogram



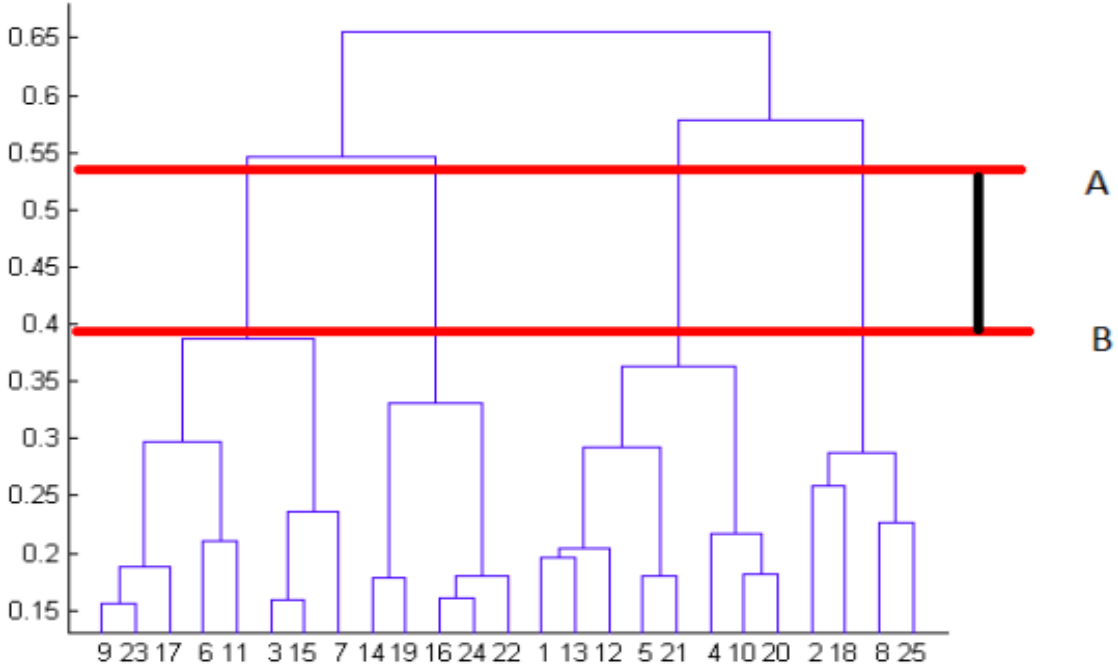
Hierarchical Clustering - Dendrogram

The height in the dendrogram at which two clusters are merged represents the distance between two clusters in the data space.

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram.

The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

Hierarchical Clustering - Dendrogram



Hierarchical Clustering

The decision of merging two clusters is taken on the basis of closeness of these clusters.

There are multiple metrics for deciding the closeness of two clusters :

- Euclidean distance
- Squared Euclidean distance
- Manhattan distance
- Maximum distance
- Mahalanobis distance

Applications

- Recommendation engines
- Market segmentation
- Social network analysis
- Search result grouping
- Medical imaging
- Anomaly detection

Quiz

Quiz

How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

1. **Creating different models for different cluster groups.**
2. **Creating an input feature for cluster ids as an ordinal variable.**
3. **Creating an input feature for cluster centroids as a continuous variable.**
4. **Creating an input feature for cluster size as a continuous variable.**

A. 1 only

B. 1 and 2

C. 3 only

D. 2 and 4

E. All of the above

Quiz

How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

1. **Creating different models for different cluster groups.**
2. **Creating an input feature for cluster ids as an ordinal variable.**
3. **Creating an input feature for cluster centroids as a continuous variable.**
4. **Creating an input feature for cluster size as a continuous variable.**

A. 1 only

B. 1 and 2

C. 3 only

D. 2 and 4

E. All of the above

Quiz

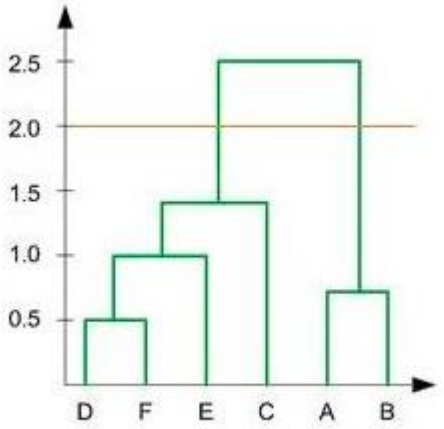
In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

A. 1

B. 2

C. 3

D. 4



Quiz

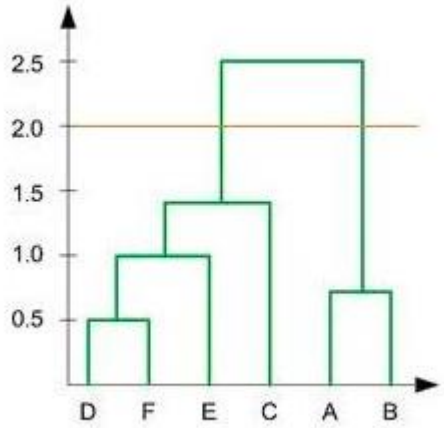
In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

A. 1

B. 2

C. 3

D. 4



Since the number of vertical lines intersecting the red horizontal line at y=2 in the dendrogram are 2, therefore, two clusters will be formed.

Quiz

In which of the following cases will K-Means clustering fail to give good results?

1. **Data points with outliers**
2. **Data points with different densities**
3. **Data points with round shapes**
4. **Data points with non-convex shapes**

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4

Quiz

In which of the following cases will K-Means clustering fail to give good results?

1. **Data points with outliers**
2. **Data points with different densities**
3. **Data points with round shapes**
4. **Data points with non-convex shapes**

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4 (*K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.)*

Quiz

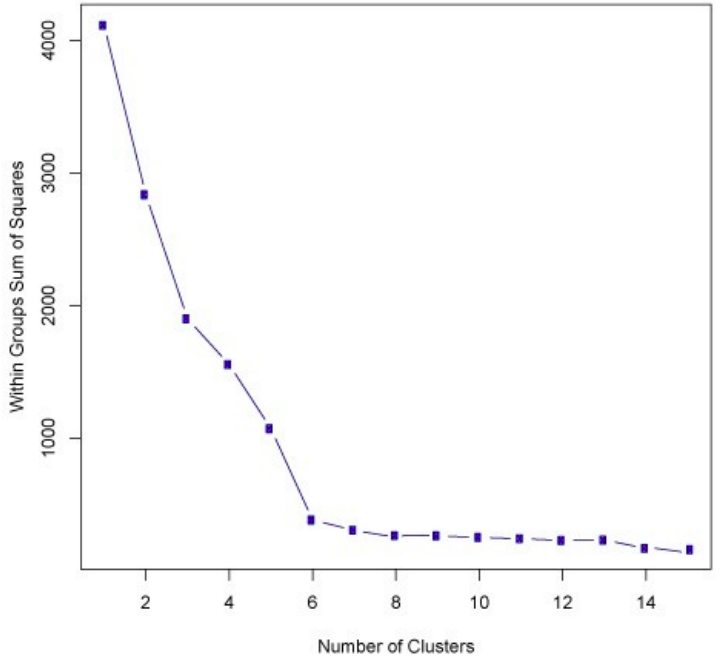
What should be the best choice for number of clusters based on the following results:

A. 5

B. 6

C. 14

D. Greater than 14



Quiz

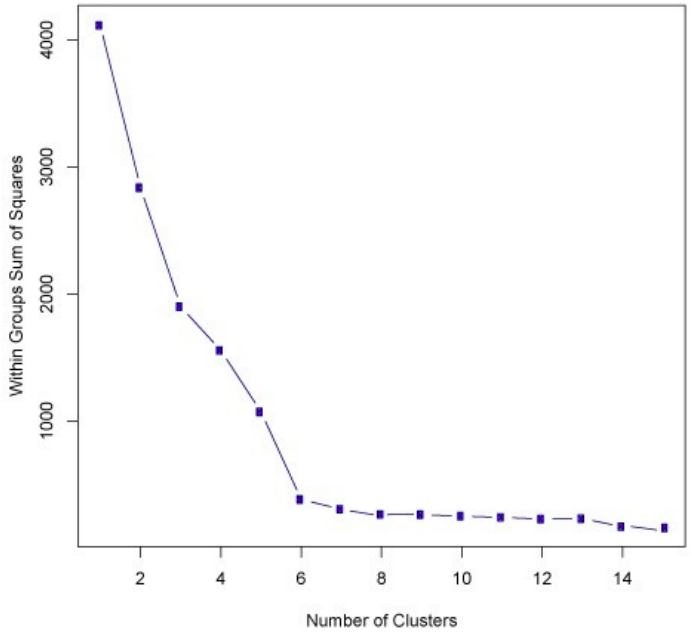
What should be the best choice for number of clusters based on the following results:

A. 5

B. 6

C. 14

D. Greater than 14



Thank you!