STA 3180 Statistical Modelling: Clustering

Clustering

Clustering is a type of unsupervised learning that is used to group data points into distinct clusters based on their similarity. It is a powerful tool for exploring and understanding data, and can be used to identify patterns and trends in data. Clustering can also be used to make predictions about future data points.

Key Concepts

- Similarity: Clustering algorithms group data points based on their similarity. This means that data points that are more similar will be grouped together, while data points that are less similar will be grouped separately.

- Clusters: Clusters are groups of data points that are similar to each other. Clusters can be identified by looking at the distance between data points.

- Distance: Distance is a measure of how far two data points are from each other. Distance is often used to determine which data points should be grouped together in a cluster.

- Centroids: A centroid is the center of a cluster. It is the point that is closest to all of the other points in the cluster.

- Density: Density is a measure of how close data points are to each other. Data points that are close together have a high density, while data points that are far apart have a low density.

Definitions

- Clustering: Clustering is a type of unsupervised learning that is used to group data points into distinct clusters based on their similarity.

- Similarity: Similarity is a measure of how similar two data points are. Data points that are more similar will be grouped together, while data points that are less similar will be grouped separately.

- Cluster: A cluster is a group of data points that are similar to each other.

- Distance: Distance is a measure of how far two data points are from each other.

- Centroid: A centroid is the center of a cluster. It is the point that is closest to all of the other points in the cluster.

- Density: Density is a measure of how close data points are to each other.

Rules

- Clustering algorithms must be able to identify clusters in data.

- The distance between data points should be used to determine which data points should be grouped together in a cluster.

- The centroid of a cluster should be the point that is closest to all of the other points in the cluster.

- Data points that are close together should have a high density, while data points that are far apart should have a low density.

Examples

- Suppose you have a dataset of customer purchases. You could use clustering to group customers based on their purchase history. Customers who have purchased similar items would be grouped together in the same cluster, while customers who have purchased different items would be grouped in different clusters.

- Suppose you have a dataset of housing prices. You could use clustering to group houses based on their price. Houses that are similar in price would be grouped together in the same cluster, while houses that are different in price would be grouped in different clusters.