STA 3180 Statistical Modelling: Clustering

Extra Practice Problems: Clustering

1. How many clusters can be formed from a set of 10 observations?

Solution: The maximum number of clusters that can be formed from a set of 10 observations is 10. This is because each observation can form its own cluster. [CORRECT]

2. How would you determine the optimal number of clusters for a given dataset?

Solution: The optimal number of clusters for a given dataset can be determined by using a clustering algorithm such as k-means or hierarchical clustering. The algorithm will iteratively group the data points into clusters and measure the quality of the clusters using a metric such as the sum of squared errors (SSE). The optimal number of clusters is the one that produces the lowest SSE. [CORRECT]

3. What is the difference between hierarchical clustering and k-means clustering?

Solution: Hierarchical clustering is an agglomerative clustering technique that starts with each data point in its own cluster and then merges clusters together based on some similarity measure. K-means clustering is a partitioning clustering technique that starts with all data points in one cluster and then iteratively assigns data points to clusters based on their distance from the cluster centroid. [CORRECT]

4. What is the formula for calculating the sum of squared errors (SSE) for a given dataset?

Solution: The formula for calculating the sum of squared errors (SSE) for a given dataset is: SSE = $\Sigma i=1n(xi - \mu)2$, where xi is the ith data point, μ is the mean of the dataset, and n is the number of data points. [CORRECT]

5. What is the formula for calculating the within-cluster sum of squares (WCSS) for a given dataset?

Solution: The formula for calculating the within-cluster sum of squares (WCSS) for a given dataset is: WCSS = $\Sigma i=1k\Sigma j=1ni=j(xi - \mu i)2$, where xi is the ith data point, μi is the mean of the ith cluster, k is the number of clusters, and n is the number of data points. [CORRECT]