STA 3100 Programming With Data in R: Clustering

# Clustering for STA 3100 Programming With Data in R

## Introduction

Clustering is a type of unsupervised learning that groups data points together into clusters based on their similarity. It is used to identify patterns and groupings in data that may not be easily visible. Clustering can be used to gain insight into the structure of a dataset, to identify outliers, and to make predictions.

## Key Concepts

- Unsupervised Learning: Unsupervised learning is a type of machine learning algorithm that does not require labeled data. The goal of unsupervised learning is to find patterns and structure in the data without any prior knowledge or labels.

- Clusters: A cluster is a group of data points that are similar to each other. Clusters are formed by grouping data points that are close together in some measure of distance.

- Distance Measure: A distance measure is a metric used to measure the similarity between two data points. Common distance measures include Euclidean distance, Manhattan distance, and cosine similarity.

- K-Means Clustering: K-means clustering is an iterative algorithm that assigns data points to clusters by minimizing the within-cluster sum of squares. It starts by randomly assigning data points to clusters and then iteratively moves them to the cluster with the closest centroid.

- Hierarchical Clustering: Hierarchical clustering is an algorithm that creates a hierarchy of clusters by merging or splitting them. It starts by assigning each data point to its own cluster and then iteratively merges or splits clusters until all data points are in the same cluster.

## Definitions

- Centroid: A centroid is the center of a cluster. It is typically calculated as the mean of all the data points in the cluster.

- Within-Cluster Sum of Squares (WCSS): WCSS is a measure of the sum of the squared distances between each data point and the centroid of its cluster.

- Dendrogram: A dendrogram is a tree-like diagram used to visualize the results of hierarchical clustering. It shows how clusters are merged or split at each step of the algorithm.

## Coding Examples

### K-Means Clustering

~~~
Start of Code
```R
~~~

```
# Load the libraries
library(tidyverse)
library(cluster)
# Load the dataset
data <- read_csv("dataset.csv")
# Perform k-means clustering
kmeans_model <- kmeans(data, centers = 3)
# Print the cluster assignments
print(kmeans_model$cluster)
```

End of Code

### Hierarchical Clustering

Start of Code

```R
# Load the libraries
library(tidyverse)
library(cluster)
# Load the dataset
data <- read_csv("dataset.csv")
# Perform hierarchical clustering
hc_model <- hclust(dist(data))
# Plot the dendrogram
plot(hc_model)
```

End of Code

## Practice Multiple Choice Questions

Q1. Which of the following is NOT a key concept of clustering?

A. Supervised learning

B. Clusters

C. Distance measure

D. Decision tree

Answer: A. Supervised learning