STA 3180 Statistical Modelling: Clustering

STA 3180 Statistical Modelling - Lecture Notes on Clustering

Introduction

Clustering is a type of unsupervised learning technique used to identify groups of similar objects within a dataset. It is an important tool for exploratory data analysis, as it can help uncover relationships between variables that may not be immediately apparent. Clustering can also be used to identify outliers in a dataset, or to create more meaningful labels for categorical variables.

Key Concepts

* **Cluster:** A group of objects that are similar to each other, but different from objects in other clusters.

* **Distance Measure:** A metric used to measure the similarity between two objects. Common distance measures include Euclidean distance and Manhattan distance.

* **Centroid:** The center of a cluster, usually defined as the mean of all points in the cluster.

* **Linkage:** The method used to determine the distance between clusters. Common linkage methods include single-linkage, complete-linkage, and average-linkage.

Algorithms

There are several algorithms used for clustering, each with its own strengths and weaknesses.

K-Means

K-Means is one of the most commonly used clustering algorithms. It works by randomly selecting k centroids, then assigning each point in the dataset to the closest centroid. The centroids are then updated to be the mean of all points assigned to them. This process is repeated until the centroids no longer change.

```
### Start of Code
# K-Means in Python
from sklearn.cluster import KMeans
# Create a KMeans model with 3 clusters
kmeans = KMeans(n_clusters=3)
# Fit the model to the data
kmeans.fit(X)
# Predict the cluster labels for new data
labels = kmeans.predict(X_new)
# End of Code
```

Hierarchical Clustering

Hierarchical clustering is a type of clustering algorithm that builds a hierarchy of clusters. It works by first assigning each point to its own cluster, then merging the closest clusters together until there is only one cluster left. The distance between clusters is determined using a linkage method.

```
### Start of Code
# Hierarchical Clustering in Python
from scipy.cluster.hierarchy import linkage, dendrogram
# Compute the linkage matrix
Z = linkage(X, 'complete')
# Generate a dendrogram
dendrogram(Z)
# End of Code
```

Practice Multiple Choice Questions

Q1. Which of the following is NOT a common distance measure used in clustering?

A. Cosine distance

B. Manhattan distance

C. Hamming distance

D. Euclidean distance

Answer: C. Hamming distance