Data Science with Python Project

Title - Hierarchical Clustering

ABSTRACT

Clustering is an unsupervised machine learning process that creates clusters such that data points inside a cluster are close to each other and also apart from data points in other clusters. There are many clustering technique to group the data objects on the basis of similarity, distance and common neighbour. The hierarchical clustering technique is one of them. This project describes the hierarchical clustering technique.

OBJECTIVE

Here in this project we are performing hierarchical clustering, where the input is a set of points, with a score that represents the pairwise similarity or dissimilarity of the points. The goal is to output a tree, often binary, whose leaves represent data points, and internal nodes represent clusters.

INTRODUCTION

Hierarchical clustering creates the hierarchical decomposition of database. The algorithm iteratively split the database into smaller subset, until some termination condition is satisfied. The hierarchical clustering algorithms do not need k as an input parameter, which is an advantage over partitioning algorithms. The hierarchical decomposition can be represented by dendrogram in two ways.
I. Bottom-up (agglomerative) approach
II. Top-down (Divisive) approach.
The basic agglomerative, hierarchical clustering algorithm works as following ways Initially each object is placed in a unique cluster. For each pair of clusters, some value of dissimilarity or distance is computed. For instance, the distance may be in minimum distances (Single linkage) in the current clustering are merged, until the whole data sets forms a single cluster.

METHODOLOGY

For performing We will use Agglomerative Clustering, a type of hierarchical clustering that follows a bottom up approach. We begin by treating each data point as its own cluster. Then, we join clusters together that have the shortest distance between them to create larger clusters. This step is repeated until one large cluster is formed containing all of the data points.

Hierarchical clustering requires us to decide on both a distance and linkage method. We will use euclidean distance and the Ward linkage method, which attempts to minimize the variance between clusters.

CODE

# Importing the libraries

```
import pandas as pd
import matplotlib.pyplot as plt
```

# Import DATASET

```
dataset = pd.read_csv('Mall_Customers.csv')
x = dataset.iloc[:,:].values
x
```

# DENDROGRAM

```
import scipy.cluster.hierarchy as sch
dendrogram = sch.dendrogram(sch.linkage(x,method ='ward'))
plt.title("Dendrogram")
plt.xlabel("Customers")
plt.ylabel("Euclidean Distance")
plt.show()
```

# Building ML Model

```
from sklearn.cluster import AgglomerativeClustering
clustering = AgglomerativeClustering(n_clusters=5)
y_hc = clustering.fit_predict(x)
y_hc
```

```
# Visualising the clusters
plt.scatter(x[y_hc ==0,0],x[y_hc == 0,1], c="red", label="C1")
plt.scatter(x[y_hc ==1,0],x[y_hc == 1,1], c="orange", label="C2")
plt.scatter(x[y_hc ==2,0],x[y_hc == 2,1], c="green", label="C3")
plt.scatter(x[y_hc ==3,0],x[y_hc == 3,1], c="blue", label="C4")
plt.scatter(x[y_hc ==4,0],x[y_hc == 4,1], c="black", label="C5")
plt.title("Cluster of Customers")
plt.xlabel("Annual Income(k$)")
plt.ylabel("Spending Score (1-100)")
plt.legend()
plt.show()
```

CONCLUSION

Based on the experimental result which we have performed on the sample Mall customers dataset we have found the relation of the customers annual income vs spending score. Hierarchical clustering is a very useful way of segmentation. The advantage of not having to pre-define the number of clusters gives it quite an edge over k-Means. However, it doesn't work well when we have huge amount of data.

Project By-

Pawan Saini