Assignment Project Report

Hierarchical K-Means: Construction of Hashing Tree

Name : Meghna Maan

Course : AI & ML (Batch 4)

Given Question

Perform Hierarchical Clustering from scratch and also using sklearn to perform wholesale customer segmentation based on their annual spending on products. You can use this dataset. Use the threshold to
1. Divide the dataset into two clusters.

2. To divide the dataset into k clusters, such that the distance between the two clusters is greater than a given threshold (this threshold can be anything passed to the function).

Prerequisites

1. Software:

    Python 3

2. Tools:

    - Numpy

    - Pandas

    - Matplotlib

    - scipy

Dataset Link: Wholesale customers data

https://archive.ics.uci.edu/ml/machine-learning- databases/00292/Wholesale%20customers%20data.csv

Methods Used

Hierarchical clustering is the hierarchical decomposition of the data based on group similarities. It allows us to build tree structures from data similarities and see how different sub- clusters relate to each other, and how far apart data points are. It gives us a tree-type structure based on the hierarchical series of nested clusters. A diagram called Dendrogram graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged, or clusters are broken apart.

Implementation

1. Loading Libraries and Dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.cluster.hierarchy as shc
%matplotlib inline
```

```
# Reading the libary
data = pd.read_csv('Wholesale customers data.csv')
data.head()
```

2. Normalizing Data

```
from sklearn.preprocessing import normalize
data_scaled = normalize(data)
data_scaled = pd.DataFrame(data_scaled, columns=data.columns)
data_scaled.head()
```

3. Sklearn

```
from sklearn.cluster import AgglomerativeClustering
cluster = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='ward')
cluster.fit_predict(data_scaled)
```

4. Plotting the final results

```
plt.figure(figsize=(10, 7))
plt.scatter(data_scaled['Milk'], data_scaled['Grocery'], c=cluster.labels_)
```

```
<matplotlib.collections.PathCollection at 0x24e62b13048>
```