Assignment Project Report

Hierarchical K-Means: Construction of Hashing Tree

Name: Arun Govind
Course: AI and ML
(Batch 4)

- **Problem Statement**

  Perform Hierarchical Clustering from scratch and also using sklearn to perform wholesale customer segmentation based on their annual spending on products. 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**

  - Software:
    - Python 3 (Use anaconda as your python distributor as well)

  - Tools:
    - Pandas
    - Numpy
    - Matplotlib

  - Dataset: Wholesale customers data from UCI archive

- **Method 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.

  Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

- **Implementation:**

1. Load all required libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

```
data = pd.read_csv('Wholesale customers data.csv')
data.head()
```

2. Scaling and preprocessing data

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

3. Plotting dendrograms

```
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(10, 7))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(data_scaled, method='ward'))
```

4. Implementing Clustering Algorithm

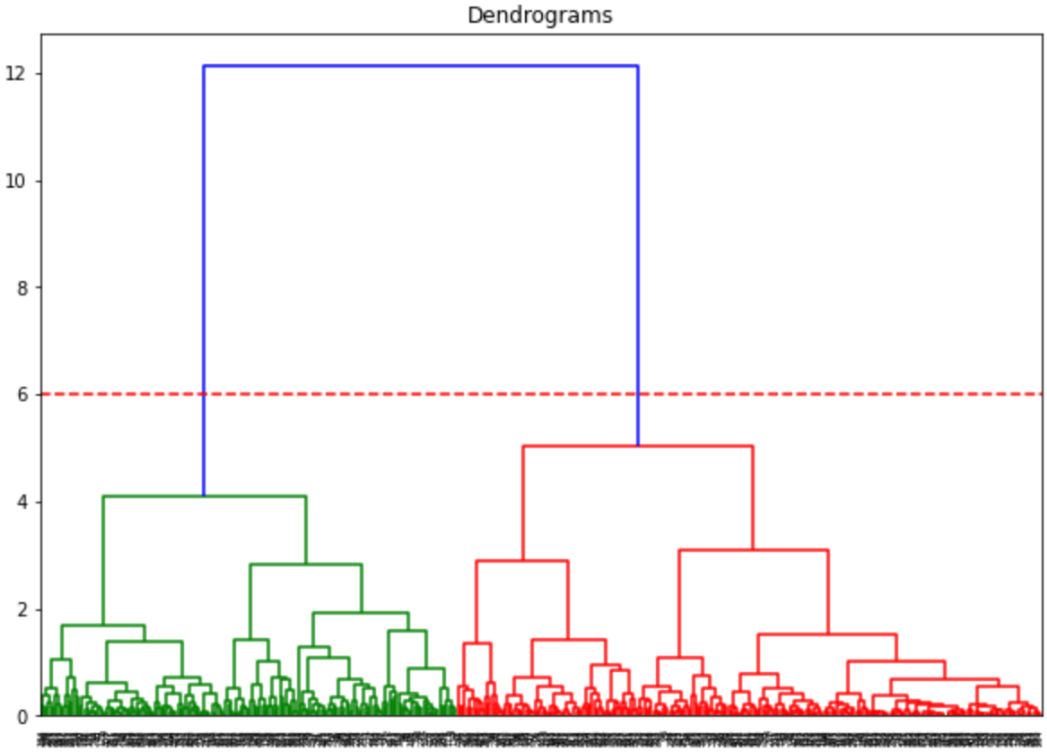
```
from sklearn.cluster import AgglomerativeClustering
cluster = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='ward')
cluster.fit_predict(data_scaled)
```

5. Plotting Graph for Clusters

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

- **Results:**

1. Plotted Dendrogram:



2. Cluster graph: