py-of-nrcm-hierarchical-clustering

August 28, 2023

0.0.1 Name : J.Keerthana

0.0.2 Roll No : 21X05A6721

0.0.3 Branch : Data Science

0.0.4 College : Narasimha Reddy Engineering College UGC Autonomous

1 Project Title:

1.0.1 Analysis and prediction of "Mall_Customers.csv" of American mall market called as phonix mall,find out on the basis of clients requirements and dendrogram using scipy graphic library with the help of "scipy.cluster.hieracy",to ace the no.of linkage of a clustering to predict.

2 Problem Statement:

2.0.1 The american finance market clients as per the rate of GDP of 2011 found as no.of growth in their business market.

2.0.2 As data science enginner find out which hierachy cluster gives maximum linkage in upcoming future

3 Task 1:

3.0.1 With the help of numpy library import the library and dataset.

4 Task 2:

4.0.1 Using the dendrogram to find the optimum number of clusters.

5 Task 3:

5.0.1 Create a hierachy model and visulaze the cluster with the cluster with the help of matplot library

6 Importing the libraries

[1]:
```
#Import the numpy, pandas , matplotlib, seaborn libery's
import numpy as np
import pandas as pd
```

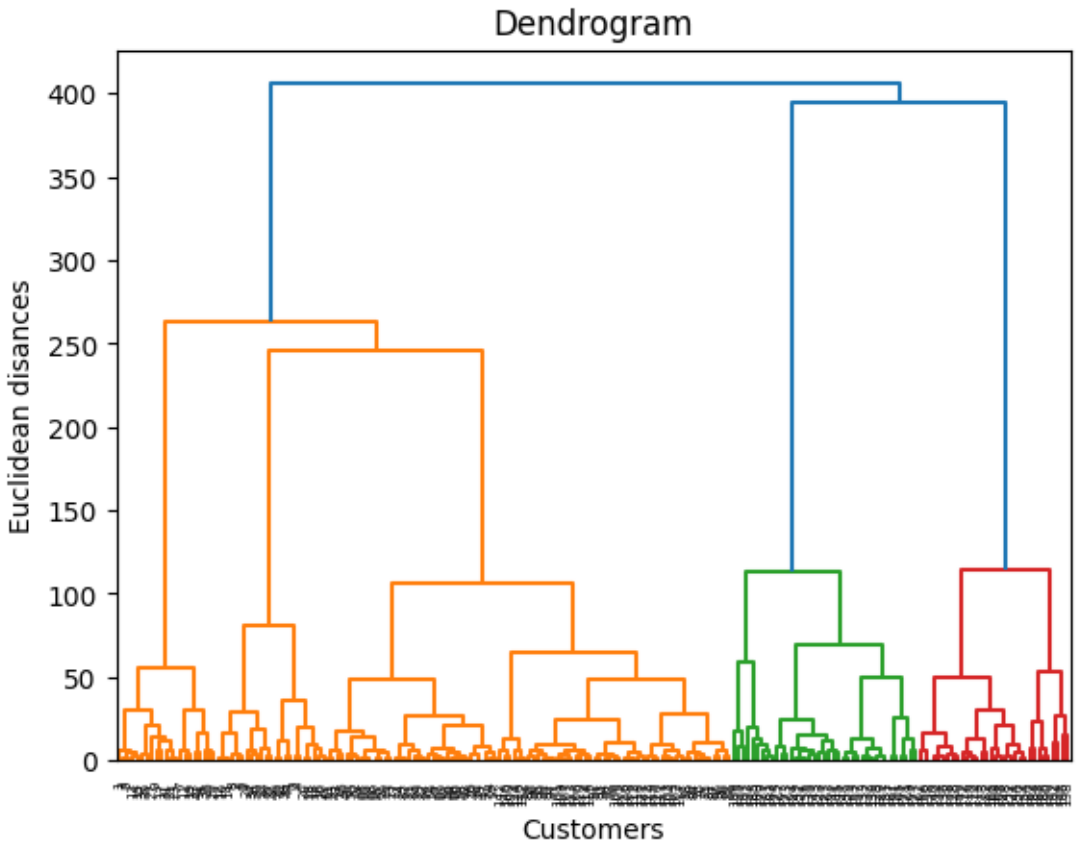
```
import matplotlib.pyplot as plt
import seaborn as sns
```

7 Loading the datasets

[4]:
```
#Assign variable name "dataset" and the input variable as "X" indcludong select␣
↪all the row and index columns which you want [colum_index, Column_index].
dataset = pd.read_csv("Mall_Customers.csv")
X = dataset.iloc[:,[3,4]].values
```

8 Using the dendrogram to find the optimal number of clusters

[5]:
```
#import scipy  cluster using attribute "scipy.cluster.hierarchy" as sch alias
import scipy.cluster.hierarchy as sch
dendrogram = sch.dendrogram(sch.linkage(X,method = 'ward'))
plt.title('Dendrogram')
plt.xlabel('Customers')
plt.ylabel('Euclidean disances')
plt.show()
```



9 Training the Hierarichal Clustering model on the dataset

[6]:
```
#Create a cluster for  five or nth cluster which you want.
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage =␣
 ↪'ward')
y_hc = hc.fit_predict(X)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_agglomerative.py:983:
FutureWarning: Attribute `affinity` was deprecated in version 1.2 and will be
removed in 1.4. Use `metric` instead
  warnings.warn(
```

10 Visualising the clusters

[8]:
```
#Plot the scatter plot for scatter  visualization.
plt.scatter(X[y_hc == 0,0], X[y_hc == 0,1 ], s = 100, c = 'red', label =␣
 ↪'Cluster 1')
plt.scatter(X[y_hc == 1,0], X[y_hc == 1,1 ], s = 100, c = 'blue', label =␣
 ↪'Cluster 2')
plt.scatter(X[y_hc == 2,0 ], X[y_hc == 2,1 ], s = 100, c = 'green', label =␣
 ↪'Cluster 3')
plt.scatter(X[y_hc == 3,0 ], X[y_hc == 3,1 ], s = 100, c = 'cyan', label =␣
 ↪'Cluster 4')
plt.scatter(X[y_hc == 4,0 ], X[y_hc == 4,1 ], s = 100, c = 'magenta', label =␣
 ↪'Cluster 5')
plt.title('Clusters of customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.legend()
plt.show()
```


Clusters of customers
Spending Score (1-100)
Annual Income (k$)
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
0
20
40
60
80
100
20
40
60
80
100
120
140

11 Conclusion:

11.1 according to the model building as enginner my prediction is cluster number 3 as give highest no.of linkage

12 Insights:

12.0.1 1.Cluster 1 contain(red) the color which shows that unsupervised learning cluster has maximum eucliding distance from the centroid up to annual income approximate 139ks.

12.0.2 2.Cluster 2 cotains(blue) which shows that unsupervised learning cluster has maximum eucliding distance from the centroid up to annual income approximately 79-80ks.

12.0.3 3.Cluster 3 cotains(green) which shows that unsupervised learning cluster has maximum eucliding distance from the centroid up to annual income approximately 139s.

12.0.4 4.Cluster 4 cotains(skyblue) which shows that unsupervised learning cluster has maximum eucliding distance from the centroid up to annual income approximately 39-40s.

12.0.5 5.Cluster 2 cotains(pink) which shows that unsupervised learning cluster has maximum eucliding distance from the centroid up to annual income approximately40-41s