Lab 6.1 Hierarchical Clustering

ລະຫັດນັກສຶກສາ: 205Q0010.19
ຊື່ ແລະ ນາມສະກຸນ: ທ້າວ ນູຊ໋ວ ເຮ 3CW1

ຈົ່ງຕອບຄຳຖາມຕໍ່ໄປນີ້ໃຫ້ສຳເລັດດ້ວຍການນຳໃຊ້ຄຳສັ່ງຂອງ Python:
ພາກທີ 1
1.1. ຈາກຊຸດຂໍ້ມູນMall_Customers.csvຈົ່ງເຕີ່ມຂໍ້ມູນໃນຕາຕະລາງໃຫ້ສຳເລັດ.

```
dataset = pd.read_csv('Mall_Customers.csv')
X = dataset.iloc[:, [3, 4]].values
print(dataset)
```

[7] ✓ 0.6s Python

	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
..	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

[200 rows x 5 columns]

1.2. ຈົ່ງທຳການເລືອກຄຸນລັກສະນະ (Features): Age ແລະ Annual Income (k$)

```
dataset = pd.read_csv('Mall_Customers.csv')
X = dataset.iloc[:, [3, 4]].values
print(dataset)
```

[17] ✓ 0.6s Python

Age	Annual Income (k$)
19	15
21	15
20	16
23	16
31	17
...	...
35	120
45	126
32	126
32	137
30	137

1.3 ຈົ່ງຊອກຫາຈຳນວນກຸ່ມເພື່ອແບ່ງກຸ່ມຊຸດຂໍ້ມູນໃຫ້ເໝາະສົມດ້ວຍDendrogram

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

[3] ✓ 4.6s Python

1.4. ຈົ່ງທຳການຝຶກຂໍ້ມູນໃນຂໍ້ທີ 1.2 ເພື່ອແບ່ງຕາມຈຳນວນກຸ່ມໃນຂໍ້ທີ 1.3 ດ້ວຍແບບຈຳລອງ (Algorithm) AgglomerativeClustering.

```
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(X)
```

[31] ✓ 0.4s Python

1.5. ຈົ່ງທຳການສຳຫຼວດ (visulaizing) ກຸ່ມຂໍ້ມູນທີ່ຖືກແບ່ງດ້ວຍscatter plot ພ້ອມອະທິບາຍຜົນ.

```
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()
```

[5] ✓ 0.2s Python

- Cluster1 (ສີແດງ): ລາຍຮັບສູງ ແລະ ຄະແນນການໃຊ້ຈ່າຍຕ່ຳ
- Cluster2 (ສີຟ້າ): ລາຍໄດ້ປົກກະຕິ ແລະ ຄະແນນການໃຊ້ຈ່າຍປົກກະຕິ
- Cluster3 (ສີຂຽວ): ລາຍຮັບສູງ ແລະ ຄະແນນການໃຊ້ຈ່າຍສູງ
- Cluster4 (ສີຟ້າຂຽວ): ລາຍຮັບຕ່ຳ ແລະ ຄະແນນການໃຊ້ຈ່າຍສູງ
- Cluster5 (ສີມ່ວງແດງ): ລາຍຮັບຕ່ຳ ແລະຄະແນນການໃຊ້ຈ່າຍຕ່ຳ

ພາກປະຕິບັດ 2

2.1 ຈົງອ່ານຊຸດຂໍ້ມູນ BaskinRoobins.csv ຈົ່ງຕື່ມຂໍ້ມູນໃນຕາຕະລາງໃຫ້ສຳເລັດ

```
dataset = pd.read_csv('BaskinRobbins.csv')
X = dataset.iloc[:, [3, 4]].values
print(dataset)
```

[21] ✓ 0.9s Python

Output exceeds the size limit. Open the full output data in a text editor

	Unnamed: 0	Flavour	Calories	Total Fat (g)	Trans Fat (g)	Carbohydrates (g)	Sugars (g)	Protein (g)
0	0	Bananas Foster	160	8.0	0.2	20	16	2.0
1	1	Baseball Nut	160	9.0	0.2	19	13	3.0
2	2	Beavertails Pastry	170	9.0	0.3	21	15	3.0
3	3	Blackberry Frozen Yogurt	120	4.0	0.1	17	16	3.0
4	4	Blue Raspberry Sherbet	130	2.0	0.1	26	20	2.0
..	...	...	...	...	...	...	...	...
65	65	Very Berry Strawberry	200	10.0	0.4	24	21	4.0
66	66	Watermelon Splash	120	0.5	0.0	27	20	0.1
67	67	Wild 'n Reckless	80	1.5	0.0	16	12	1.0
68	68	Winter White Chocolate	160	8.0	0.2	20	16	2.0
69	69	World Class Chocolate	260	60.0	0.4	25	18	5.0

67 NaN
68 NaN
69 NaN

[70 rows x 9 columns]

2.2. ຈົ່ງທຳການເລືອກຄຸນລັກສະນະ (Features): Carbohydrates (g) ແລະ Sugars (g)

```
dataset = pd.read_csv('BaskinRobbins.csv')
X = dataset.iloc[:, [3, 4]].values
print(dataset)
```

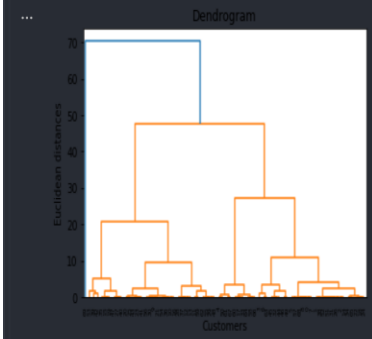
[19] ✓ 0.1s Python

Carbohydrates (g)	Sugars (g)
20	16
19	13
21	15
17	16
26	20
...	...
24	21
27	20
16	12
20	16
25	18

2.3. ຈົ່ງຊອກຫາຈຳນວນກຸ່ມເພື່ອແບ່ງກຸ່ມຊຸດຂໍ້ມູນໃຫ້ເໝາະສົມດ້ວຍDendrogram

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

[10] ✓ 0.9s Python



2.4. ຈົ່ງທຳການຝຶກຂໍ້ມູນໃນຂໍ້ທີ 2.2 ເພື່ອແບ່ງຕາມຈຳນວນກຸ່ມໃນຂໍ້ທີ 2.3 ດ້ວຍແບບຈຳຮອງ (Algorithm) AgglomerativeClustering.

```
from sklearn.cluster import AgglomerativeClustering
hc = AgglomerativeClustering(n_clusters = 5, affinity = 'euclidean', linkage = 'ward')
y_hc = hc.fit_predict(X)
```

[31] ✓ 0.4s Python

2.5. ຈົ່ງທຳການສຳຫຼວດ (visulaizing) ກຸ່ມຂໍ້ມູນທີ່ຖືກແບ່ງດ້ວຍ scatter plot ພ້ອມອະທິບາຍຜົນ.

❖

```
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()
```

[12] ✓ 0.2s Python

- ❖ Cluster1 (ສີແດງ): ລາຍຮັບຕ່ຳ ແລະ ຄະແນນການໃຊ້ຈ່າຍບາງເທື່ອກ່ຳສູງບາງເທື່ອກ່ຳຕ່ຳ
- ❖ Cluster2 (ສີຟ້າ): ລາຍໄດ້ປົກກະຕິ ແລະ ຄະແນນການໃຊ້ຈ່າຍສູງ
- ❖ Cluster3 (ສີຂຽວ): ລາຍຮັບຕ່ຳ ແລະ ຄະແນນການໃຊ້ຈ່າຍສູງ
- ❖ Cluster4 (ສີຟ້າຂຽວ): ລາຍຮັບຕ່ຳ ແລະ ຄະແນນການໃຊ້ຈ່າຍສູງ
- ❖ Cluster5 (ສີມ່ວງແດງ): ລາຍຮັບສູງ ແລະຄະແນນການໃຊ້ຈ່າຍສູງ