6509611684 ณัชชา สุขะเมือง

Homework: การจัดกลุ่ม (Clustering)

1. ต้องการจัดกลุ่มของกลุ่มเมฆจากภาพถ่ายดาวเทียมแผนที่สภาพความกดอากาศ เพื่อจะนับจำนวนกลุ่มเมฆ สำหรับใช้เป็นข้อมูลในการพยากรณ์การเกิดพายุ หากกลุ่มเมฆมีการเรียงตัวติดกันมากกว่าเท่ากับ 3 ช่องนับเป็นพายุ 1 ลูก จงใช้วิธีการจัดกลุ่ม DBSCAN จับกลุ่มพายุดังกล่าว โดยให้ ‘H’ แทน บริเวณความกดอากาศสูงเป็นตัวอย่างใดใด จำนวนช่องว่างระหว่าง ‘H’ แทนระยะห่างระหว่างตัวอย่าง กำหนดให้ค่ารัศมีเท่ากับ 1 หน่วย (นับตามแนวนอนและแนวตั้ง ไม่นับตามแนวทแยงมุม) และจำนวนตัวอย่างต่ำสุดเท่ากับ 3 จงแสดงวิธีทำตามขั้นตอนวิธี DBSCAN และ ระบุด้วยว่าจับกลุ่มได้จำนวนพายุกี่ลูก

	1	2	3	4	5	6	7	8	9	10
1	H	H	H	H				H		
2		H				H		H		
3	H	H			H	H	H	H	H	H
4						H			H	
5		H		H						
6		H		H						
7	H	H				H	H		H	
8		H		H						
9	H	H						H		
10		H			H					H

กำหนดค่า

- รัศมี (ε) = 1 หน่วย
- จำนวนจุดขั้นต่ำ (MinPts) = 3 จุด

กลุ่มที่ 1

สุ่มจุดที่ (1,1) หาจุดในรัศมี แล้ววนซ้ำทุกจุดในกลุ่ม

-> set = {(1,1),(1,2),(1,3),(1,4),(2,2),(2,3),(2,4)}

ดังนั้น คลัสเตอร์กลุ่มที่ 1 ได้แก่ {(1,1),(1,2),(1,3),(1,4),(2,2),(2,3),(2,4)}

กลุ่มที่ 2

สุ่มจุดที่ (1,8) หาจุดในรัศมี แล้ววนซ้ำทุกจุดในกลุ่ม

-> set = {(1,8),(2,6),(2,8),(3,5),(3,6),(3,7),(3,8),(3,9),(3,10),(4,6),(4,9)}

ดังนั้น คลัสเตอร์กลุ่มที่ 2 ได้แก่ {(1,8),(2,6),(2,8),(3,5),(3,6),(3,7),(3,8),(3,9),(3,10),(4,6),(4,9)}

กลุ่มที่ 3

สุ่มจุดที่ (5,2) หาจุดในรัศมี แล้ววนซ้ำทุกจุดในกลุ่ม

-> set = {(5,2),(6,2),(7,1),(7,2),(8,2),(9,1),(9,2),(10,2)}

ดังนั้น คลัสเตอร์กลุ่มที่ 3 ได้แก่ {(5,2),(6,2),(7,1),(7,2),(8,2),(9,1),(9,2),(10,2)}

Noise ประกอบไปด้วย

set = {(5,4),(6,4),(8,4),(7,6),(7,7),(9,8),(7,9),(10,6),(10,10)}

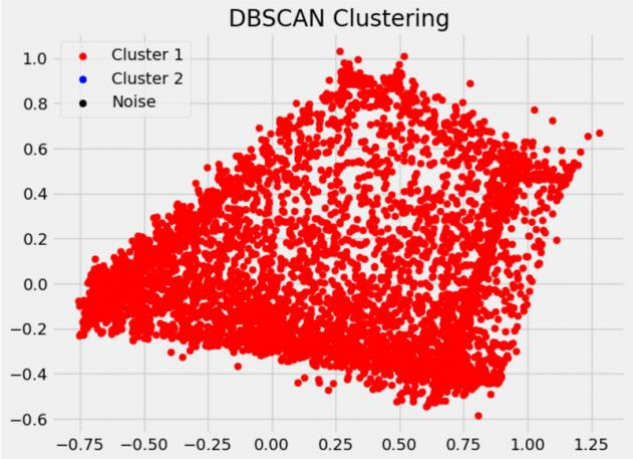
ตอบ ดังนั้น มีจำนวนพายุ 3 กลุ่ม

2. จากข้อมูลเครดิตของผู้ใช้จำนวน 8,950 คน ตาม URL: https://www.kaggle.com/arjunbhasin2013/ccdata จงใช้เครื่องมือ Clustering จับกลุ่มตามพฤติกรรมการใช้งาน เพื่อตอบคำถามต่อไปนี้

2.1. ควรใช้ขั้นตอนวิธีในการจัดกลุ่มแบบ k-means หรือ DBSCAN เพราะเหตุใด

ตอบ อ้างอิงจากโค้ดที่อาจารย์ให้มาซึ่งเป็น K-means Clustering จะเห็นได้ว่า K-means Clustering ไม่เหมาะกับการจัดกลุ่มชุดข้อมูลนี้ โดยการจัดกลุ่มข้อมูลชุดนี้ควรจัดกลุ่มแบบ DBSCAN เพราะว่าข้อมูลมีรูปร่างของกลุ่มไม่เป็นทรงกลม และมีความหนาแน่นไม่สม่ำเสมอ โดย DBSCAN จะจัดกลุ่มจากความหนาแน่นของจุดข้อมูล ซึ่งจะเหมาะกับข้อมูลชุดนี้

นี่คือ การจัดกลุ่ม DBSCAN ที่มี eps=3, min_samples=6



โดยเราทราบค่า eps และ min_sample ที่เหมาะสมจากโค้ดนี้

```
# DBSCAN Model
from sklearn.cluster import DBSCAN
from sklearn.decomposition import PCA

# ทดลองปรับค่า eps และ min_samples
eps_values = [1.5, 2, 2.5, 3]
min_samples_values = [4, 5, 6]

best_clusters = 0
```

```
best_eps = 0
best_min_samples = 0

for eps in eps_values:
    for min_samples in min_samples_values:
        dbscan = DBSCAN(eps=eps, min_samples=min_samples)
        dbscan.fit(x)

        y = dbscan.labels_
        num_clusters = len(set(y)) - (1 if -1 in y else 0)  # จำนวน Cluster ที่ไม่รวม noise
        num_noise = list(y).count(-1)  # จำนวน Noise

        if num_clusters == 1 and num_noise == 0:  # ถ้าจำนวน Cluster = 1 และไม่มี Noise
            best_clusters = num_clusters
            best_eps = eps
            best_min_samples = min_samples
            print(f"Found optimal eps = {eps}, min_samples = {min_samples}")

# ใช้ค่า eps และ min_samples ที่ดีที่สุด
print(f"Best eps: {best_eps}, Best min_samples: {best_min_samples}")

# สร้าง DBSCAN Model ด้วยค่า eps และ min_samples ที่ดีที่สุด
dbscan = DBSCAN(eps=best_eps, min_samples=best_min_samples)
dbscan.fit(x)

y = dbscan.labels_
data_with_clusters = data.copy()  # data คือ DataFrame ที่คุณใช้
data_with_clusters['Clusters'] = y

print('จำนวน Cluster (ไม่รวม noise):', len(set(y)) - (1 if -1 in y else 0))
print('จำนวน Noise (label -1):', list(y).count(-1))
```

โดยโค้ดนี้ ทดลองหลายๆค่าของ eps และ min_samples เพื่อหาค่าที่เหมาะสมที่สุดในการใช้กับโมเดล DBSCAN โดย วนลูปค่าต่างๆ ของ eps และ min_samples

2.2. แสดงตัวอย่างข้อมูลหลังผ่านกระบวนการ Normalization และ Outlier detection

ตอบ

- ข้อมูลหลังผ่านกระบวนการ Normalization

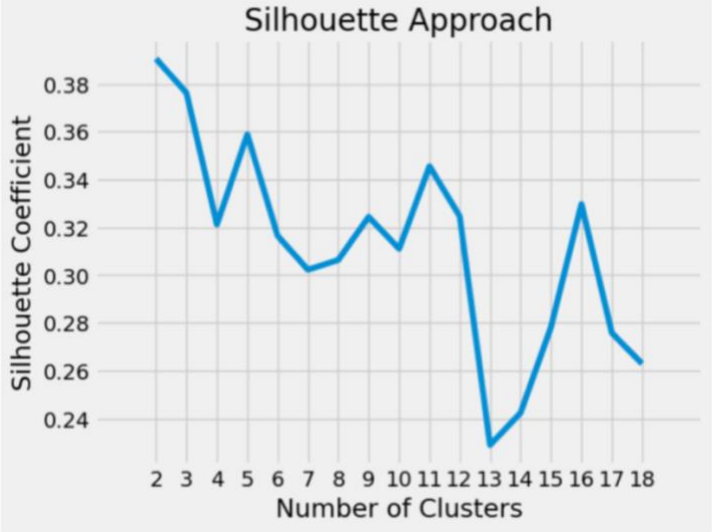
```
[[2.14779454e-03 8.18182000e-01 1.94536779e-03 ... 1.82564563e-03
  0.00000000e+00 1.00000000e+00]
 [1.68169097e-01 9.09091000e-01 0.00000000e+00 ... 1.40344791e-02
  2.22222000e-01 1.00000000e+00]
 [1.31026136e-01 1.00000000e+00 1.57662475e-02 ... 8.20961806e-03
  0.00000000e+00 1.00000000e+00]
 ...
 [1.22871936e-03 8.33333000e-01 2.94456089e-03 ... 1.07843629e-03
  2.50000000e-01 0.00000000e+00]
 [7.06688341e-04 8.33333000e-01 0.00000000e+00 ... 7.29475795e-04
  2.50000000e-01 0.00000000e+00]
 [1.95717777e-02 6.66667000e-01 2.22932216e-02 ... 1.15527021e-03
  0.00000000e+00 0.00000000e+00]]
```

- ข้อมูลหลังผ่านกระบวนการ Outlier detection

```
จำนวนข้อมูลก่อนลบ Outliers (Z-Score): 8636
จำนวนข้อมูลหลังลบ Outliers (Z-Score): 7190
```

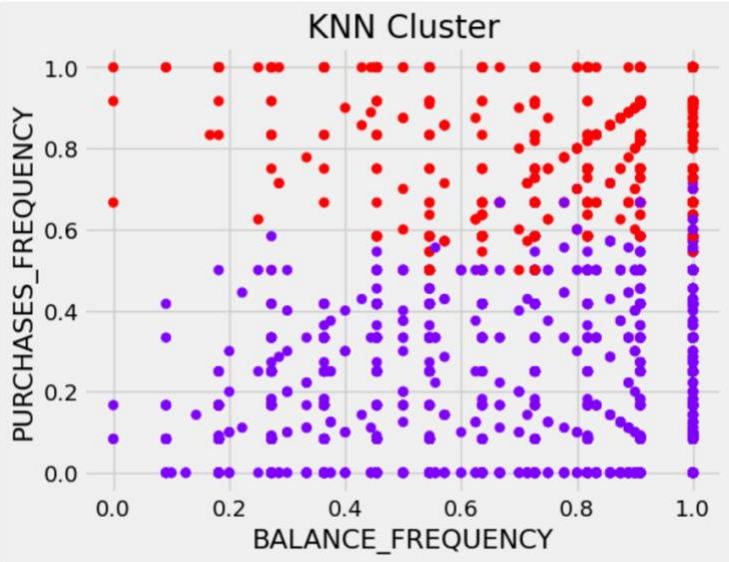
2.3. ควรแบ่งข้อมูลออกเป็นกี่กลุ่ม เพราะอะไร (การตัดสินใจควรอ้างอิงกับค่าประสิทธิภาพ แบบ Intrinsic Method โดยใช้แนวทางการพิจารณาจากค่าประสิทธิภาพต่าง ๆ เช่น Silhouette coefficient)

ตอบ 2 กลุ่ม เพราะอ้างอิงจาก Silhouette coefficient ที่มีค่า score number of 2 cluster สูงสุดเลย



2.4. อธิบายลักษณะเฉพาะของแต่ละกลุ่มว่าแตกต่างจากกลุ่มอื่นอย่างไร (อ้างอิงจากตัวแทนของกลุ่ม หรือการกระจายของกลุ่ม)

ตอบ อ้างอิงจากกราฟแสดงข้อมูล 2 กลุ่มที่ถูกจัดกลุ่มด้วย K-Means โดยแยกตามสีม่วง และ แดง



กลุ่มที่ 1 (จุดสีแดง) : Active Spenders

ลักษณะ:

- PURCHASES_FREQUENCY **สูง**: กลุ่มนี้มีความถี่ในการซื้อสูงเกือบถึง 1 (หมายถึงมีการซื้อเกือบทุกเดือน)
- BALANCE_FREQUENCY **ปานกลางถึงสูง**: หลายจุดอยู่ในช่วง 0.5 - 1 แสดงว่าลูกค้าในกลุ่มนี้มีการรักษายอดเงินในบัญชีอย่างสม่ำเสมอ

การตีความ:

- เป็นกลุ่มลูกค้าที่ **มีการใช้จ่ายบ่อย** และ **รักษายอดเงินในบัญชีสม่ำเสมอ**
- อาจเป็นกลุ่มลูกค้าที่มีการวางแผนการเงินดี หรือใช้บัตรเครดิตเป็นประจำเพื่อคะแนนสะสมหรือสิทธิประโยชน์
- พวกเขาใช้บัตรเครดิตบ่อย และมีแนวโน้มจะรักษายอดคงเหลือบ่อยครั้ง

กลุ่มที่ 2 (จุดสีม่วง) : Occasional Users

ลักษณะ:

- PURCHASES_FREQUENCY **ต่ำ**: ส่วนใหญ่ต่ำกว่า 0.5 บางจุดใกล้ 0
- BALANCE_FREQUENCY **หลากหลาย**: กระจายตั้งแต่ 0 จนถึง 1 แต่หลายจุดอยู่ด้านล่าง

การตีความ:

- เป็นกลุ่มลูกค้าที่ **ไม่ได้ใช้จ่ายบ่อย** และ **บางรายอาจไม่ค่อยรักษายอดเงิน**
- อาจเป็นกลุ่มที่ใช้บัตรเครดิตเฉพาะบางกรณีเท่านั้น หรือเป็นผู้ใช้ใหม่/ไม่ค่อยใช้งาน
- ลูกค้าในกลุ่มนี้มีการใช้งานน้อย อาจใช้บัตรในสถานการณ์เฉพาะ