**การทดลองจัดกลุ่มข้อมูล 2 มิติ ด้วยวิธี k-Means และ Hierarchical Clustering**

**ชุดข้อมูลที่ 1**

1. อ่านข้อมูล

data = pd.read\_csv('data2Dset1.csv', header=None)

1. Plot จุดข้อมูล

x = data[0]

y = data[1]

plt.scatter(x,y)

plt.show()

Chart, scatter chart

Description automatically generated

1. จัดกลุ่มข้อมูลด้วย K-Means
   1. k=2

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=2*

* 1. k=3

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=3*

* 1. k=4

A screenshot of a computer

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

*Scatter plot with k=4*

โดยสรุปจาก k ทั้ง 3 ค่าที่ได้ทดลองใช้ พบว่าค่า k ที่ดีที่สุดคือ 3 เพราะจากข้อมูลที่ได้มาเมื่อนำมา plot แล้วจะเห็นได้ว่าข้อมูลมีการแบ่งออกเป็น 3 กลุ่มชัดเจน และเมื่อใช้สีในการแยก cluster ก็จะเห็นได้ว่าข้อมูลที่อยู่ในกราฟถูกแย่งแยกอย่างชัดเจนมากขึ้น

1. จัดกลุ่มข้อมูลด้วย Hierarchical Clustering
   1. Method single

Graphical user interface, text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

*Cutoff โดยกำหนด t=1.5*

* 1. Method complete

Text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff โดยกำหนด t=6*

* 1. Method centroid

Text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff โดยกำหนด t=4*

**ชุดข้อมูลที่ 2**

1. อ่านข้อมูล

data = pd.read\_csv('data2Dset2.csv', header=None)

1. Plot จุดข้อมูล

x = data[0]

y = data[1]

plt.scatter(x,y)

plt.show()

Chart, scatter chart

Description automatically generated

1. จัดกลุ่มข้อมูลด้วย K-Means
   1. k=10

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=10*

* 1. k=4

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=4*

* 1. k=12

Text

Description automatically generated

Chart

Description automatically generated

*Scatter plot with k=12*

โดยสรุปแล้ว k ที่ทดลองนำมาใช้แล้วคิดว่าดีที่สุดคือ k=10 เนื่องจากหากมองที่ k=4 จะเห็นว่าถึงแม้จะสามารถแยกข้อมูลตามสีได้ชัด แต่กลุ่มแต่ละกลุ่มนั้นกว้างเกินไป และหากเป็น k=12 การแบ่งกลุ่มนั้นละเอียดเกินไปจนมองยาก ดังนั้น k=10 จึงเป็นตัวเลอกที่ดีที่สุดเพราะนอกจากจะแยกข้อมูลได้ไม่กว้างเกินไปแล้วก็ยังไม่ละเอียดเกินไปอีกด้วย

1. จัดกลุ่มข้อมูลด้วย Hierarchical Clustering
   1. Method single

Graphical user interface, text

Description automatically generated

Chart

Description automatically generated

Text

Description automatically generated with medium confidence

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=1*

* 1. Method ward

Text

Description automatically generated

Chart, histogram, box and whisker chart

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=2.5*

* 1. Method centroid

Text

Description automatically generated

Chart

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=1*

**ชุดข้อมูลที่ 3**

1. อ่านข้อมูล

data = pd.read\_csv('data2Dset3.csv', header=None)

1. Plot จุดข้อมูล

x = data[0]

y = data[1]

plt.scatter(x,y)

plt.show()

Chart, scatter chart

Description automatically generated

1. จัดกลุ่มข้อมูลด้วย K-Means
   1. k=5

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=5*

* 1. k=6

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=6*

* 1. k=8

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=8*

โดยสรุปแล้ว k ที่ดีที่สุดคือ k=6 ถึงแม้ข้อมูลที่มีจะดูกระจัดกระจาย แต่เมื่อลองใช้ k=6 แล้วจะเห็นข้อมูลที่จัดกลุ่มไว้ได้ชัดเจนและไม่กว้างไม่ละเอียดเกินไป หากเลข k มากกว่านี้จะได้ข้อมูลที่อ่านและวิเคราะห์ได้ยาก

1. จัดกลุ่มข้อมูลด้วย Hierarchical Clustering
   1. Method single

Graphical user interface, text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=0.2*

* 1. Method ward

Graphical user interface, text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Text

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=2.5*

* 1. Method average

A screenshot of a computer

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

Text

Description automatically generated with low confidence

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=0.5*

**ชุดข้อมูลที่ 4**

1. อ่านข้อมูล

data = pd.read\_csv('data2Dset4.csv', header=None)

1. Plot จุดข้อมูล

x = data[0]

y = data[1]

plt.scatter(x,y)

plt.show()

Chart, scatter chart

Description automatically generated

1. จัดกลุ่มข้อมูลด้วย K-Means
   1. k=5

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=5*

* 1. k=8

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=8*

* 1. k=4

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Scatter plot with k=4*

โดยสรุป k ที่ดีที่สุดคือ 4 เพราะข้อมูลแบ่งสัดส่วนกันสวยงามชัดเจน ดูง่าย แต่ถ้าหากมองว่าชุดข้อมูลกว้างเกินไปก็สามารถเลือก k=5 ได้เช่นกัน

1. จัดกลุ่มข้อมูลด้วย Hierarchical Clustering
   1. Method single

Graphical user interface, text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=0.4*

* 1. Method ward

Text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=2*

* 1. Method average

Text

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

*Cutoff ที่ t=1*

**สรุปผลการทดลอง**

K-means นั้นเป็นการระบุจำนวน cluster ที่เราต้องการจะแยกหมวดหมู่ ซึ่งวิธีนี้ควรใช้กับชุดข้อมูลที่เราเห็น pattern ชัดเจน สามารถระบุจำนวน cluster หรือ k ได้ง่าย ซึ่งการที่เราสามารถระบุ k ที่เราต้องการลงไปได้เลยนั้นก็คือข้อดีของ K-means แต่ข้อเสียคือค่อนข้างเสีย performance ในการทำงาน ใช้เวลามาก

ส่วน Hierarchical Clustering จากการทดลอง จะเป็นการลองใช้ method ทีละแบบ ซึ่งแต่ละแบบอาจจะได้ผลลัพธ์หรือการทำ dendrogram ได้ผลต่างหรือใกล้เคียงกันขึ้นอยู่กับข้อมูล แต่ทุกแบบจะทำให้เราเห็น hierarchy ที่ชัดเจนเห็น pattern ได้ดีขึ้น และเมื่อเราเห็นข้อมูลที่ถูกจัดอยู่ในรูปแบบของ dendrogram แล้ว เราก็สามารถที่จะวิเคราะห์เพื่อเลือกค่า t ที่จะนำไปใช้ cutoff เพื่อทำให้เกิด cluster ที่เหมาะสมได้ ซึ่งหมายความว่าวิธีนี้จะใช้ได้ดีก็ต่อเมื่อข้อมูลมีจำนวนไม่เยอะมาก หากข้อมูลมีเยอะเกินไป จะทำให้วิเคราะห์ค่า t ที่เหมาะสมได้ยาก