3 การวิเคราะห์เชิงพรรณนา และการวิเคราะห์เชิงวินิจฉัย

การสำรวจข้อมูลที่ได้ถูกรวบรวมไว้ นับเป็นก้าวแรกของการวิเคราะห์ข้อมูล ทั้งนี้เพื่อทำให้เกิดความเข้าใจในลักษณะของข้อมูลที่มีอยู่ โดยส่วนใหญ่มักเป็นการสังเกตลักษณะการกระจายของค่าข้อมูลในแต่ละตัวแปร ซึ่งจัดเป็นการวิเคราะห์ข้อมูลในระดับ การวิเคราะห์เชิงพรรณนา รวมไปถึงการอธิบายความสัมพันธ์ระหว่างตัวแปรในชุดข้อมูล ซึ่งจัดเป็นการวิเคราะห์เชิงวินิจฉัย แม้ว่าทั้งการวิเคราะห์เชิงพรรณนาและการวิเคราะห์เชิงวินิจฉัย จะเป็นการวิเคราะห์ข้อมูลเบื้องต้นที่มีวิธีการวิเคราะห์ข้อมูลที่ซับซ้อนไม่มากนัก แต่สามารถนำไปสู่คำตอบของปัญหาธุรกิจบางปัญหาได้ อีกทั้งผลลัพธ์จากการวิเคราะห์ข้อมูลในระดับขั้นนี้ยังสามารถนำไปใช้เป็นแนวทางในการตั้งสมมติฐานและการดำเนินการในการวิเคราะห์ข้อมูลในระดับขั้นที่สูงขึ้นไป ในบทนี้จะกล่าวถึงเครื่องมือทางคณิตศาสตร์และสถิติอย่างง่าย สำหรับใช้ในการวิเคราะห์เชิงพรรณนาและการวิเคราะห์เชิงวินิจฉัย เพื่อเป็นแนวทางในการดำเนินการวิเคราะห์ข้อมูลและนำไปต่อยอดความรู้ของผู้อ่านต่อไปในอนาคต

3.1 สถิติศาสตร์เชิงพรรณนาด้วยตารางตัวหลัก

พิจารณาตารางข้อมูล ซึ่งประกอบด้วยคอลัมน์และแถวข้อมูล โดยแต่ละแถวข้อมูล บันทึก คุณลักษณะ ที่ สนใจ ตาม คอลัมน์ ของ ข้อมูล แต่ละ ข้อมูล (แสดงตัวอย่าง ตารางข้อมูล ดังตาราง 2.1) เมื่อนำเอาคอลัมน์ใดคอลัมน์หนึ่งของตารางข้อมูล ซึ่งหมายถึง ค่าของตัวแปรใดตัวแปรหนึ่ง ของทุกๆ ข้อมูล มาพิจารณาแล้ว เราสามารถอธิบายลักษณะการกระจายของค่าตัวแปรนั้นๆ ได้หลายวิธี วิธีหนึ่งในนั้นคือการอธิบายโดยใช้ค่าสถิติเชิงพรรณนา

ค่ากลางของข้อมูล

ค่ากลางข้อมูล เป็นค่าสถิติที่ใช้ในการบ่งบอกถึงแนวโน้มของค่าข้อมูลส่วนใหญ่ ซึ่งสามารถใช้เป็นตัวแทนค่าข้อมูลได้ เราสามารถหาค่ากลางข้อมูลได้ 3 แบบ คือ

ค่าเฉลี่ยเลขคณิต (Arithmetic Mean) เรียกอย่างย่อว่า ค่าเฉลี่ย (Mean หรือ Average) คือ ค่าผลบวกของค่าข้อมูลทั้งหมดหารด้วยจำนวนข้อมูล กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, ..., x_{n,j})$ เป็นชุดของค่าตัวแปร X_j ของทุกๆ ข้อมูลในตารางข้อมูล เราสามารถคำนวณค่าเฉลี่ยของตัวแปร X_j (แทนด้วยสัญลักษณ์ $\overline{x}_j$ เมื่อค่าตัวแปร X_j มาจากกลุ่มตัวอย่าง หรือ μ_j เมื่อค่าตัวแปร X_j มาจากประชากร) ได้จากสูตร

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \qquad (3.1)$$

เนื่องจากการคำนวณค่าเฉลี่ยอยู่บนพื้นฐานการดำเนินการเชิงตัวเลข ดังนั้น ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงปริมาณเท่านั้น ที่จะสามารถคำนวณหาค่าเฉลี่ยได้

ตัวอย่าง 3.1

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{\text{JP}} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัท จำนวน 10 คน จะสามารถหาค่าเฉลี่ย ได้ดังนี้

$$\overline{x}_{\text{JP}} = \frac{1}{n}\sum_{i=1}^{n} x_{i,\text{JP}}$$
$$= \frac{7 + 10 + 11 + 15 + 10 + 10 + 12 + 14 + 16 + 12}{10}$$
$$= \frac{117}{10} = 11.7$$

ดังนั้น ค่าเฉลี่ยของตัวแปรประสิทธิภาพของงาน ($\overline{x}_{\text{JP}}$) มีค่าเท่ากับ 11.7

อภิธานศัพท์

ประชากร (Population)
หมายถึง กลุ่ม สมาชิก ทั้งหมด ของ สิ่งต่างๆ ที่ต้องการศึกษาหรือต้องการสรุปอ้างอิงจะเป็น คน สัตว์ สิ่งของ เหตุการณ์ ปรากฏการณ์ หรือพฤติกรรมใดๆ ก็ได้ ซึ่งอยู่ภายในขอบเขตที่กำหนด

ตัวอย่าง (Sample)
หมายถึง กลุ่ม สมาชิก ของ ประชากร ที่ถูกเลือกมาด้วยวิธีการต่างๆ เพื่อทำการศึกษาวิเคราะห์ แล้วนำผลหรือข้อสรุปที่ได้ไปใช้อ้างอิงถึงประชากร ถ้ากลุ่มสมาชิกที่เลือกมานั้นเป็นตัวแทนที่ดีของประชากร และมีจำนวนมากพอแล้วค่าที่ใช้สรุปอ้างอิงถึงประชากรจะมีความถูกต้อง หรือใกล้เคียงกับลักษณะหรือคุณสมบัติของประชากรมาก

มัธยฐาน (Median) คือ ค่าข้อมูลตรงกลางของข้อมูลทั้งหมดซึ่งถูกเรียงลำดับจากค่าน้อยไปค่ามาก หรือจากค่ามากไปค่าน้อยแล้ว การหาค่ามัธยฐานจะต้องทำการเรียงลำดับของข้อมูลก่อนเสมอ ต่อมาจึงหาตำแหน่งตรงกลาง ซึ่งสามารถแบ่งได้ 2 กรณี ดังนี้

กรณีที่ 1 จำนวนข้อมูล n เป็นเลขคี่ สามารถหาค่ามัธยฐานของตัวแปร X_j แทนด้วยสัญลักษณ์ $\tilde{x}_j$ ได้โดย

$$\tilde{x}_j = x_{\frac{(n+1)}{2},j} \qquad (3.2)$$

กรณีที่ 2 จำนวนข้อมูล n เป็นเลขคู่ สามารถหาค่ามัธยฐานของตัวแปร X_j ได้โดย

$$\tilde{x}_j = \frac{x_{\frac{n}{2},j} + x_{\frac{n}{2}+1,j}}{2} \qquad (3.3)$$

เนื่องจากการหาค่ามัธยฐานจะต้องทำการเรียงลำดับข้อมูล ดังนั้น ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงลำดับ หรือ ข้อมูลเชิงปริมาณ เท่านั้น ที่จะสามารถคำนวณค่ามัธยฐานได้

ตัวอย่าง 3.2

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{\text{JP}} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัทจำนวน 10 คน จะสามารถหาค่ามัธยฐานได้โดยเรียงลำดับข้อมูลจากค่าน้อยไปค่ามาก ได้ดังนี้ $(7, 10, 10, 10, 11, 12, 12, 14, 15, 16)$ ชุดค่าข้อมูลนี้มีจำนวนข้อมูล n เท่ากับ 10 ซึ่งเป็นเลขคู่ ดังนั้น ค่ามัธยฐาน สามารถคำนวณได้โดยใช้สมการ (3.3) ดังนี้

$$\begin{aligned}
\tilde{x}_{\text{JP}} &= \frac{x_{\frac{n}{2},\text{JP}} + x_{\frac{n}{2}+1,\text{JP}}}{2} \\
&= \frac{x_{\frac{10}{2},\text{JP}} + x_{\frac{10}{2}+1,\text{JP}}}{2} \\
&= \frac{x_{5,\text{JP}} + x_{6,\text{JP}}}{2} \\
&= \frac{11 + 12}{2} \\
&= \frac{23}{2} = 11.5
\end{aligned}$$

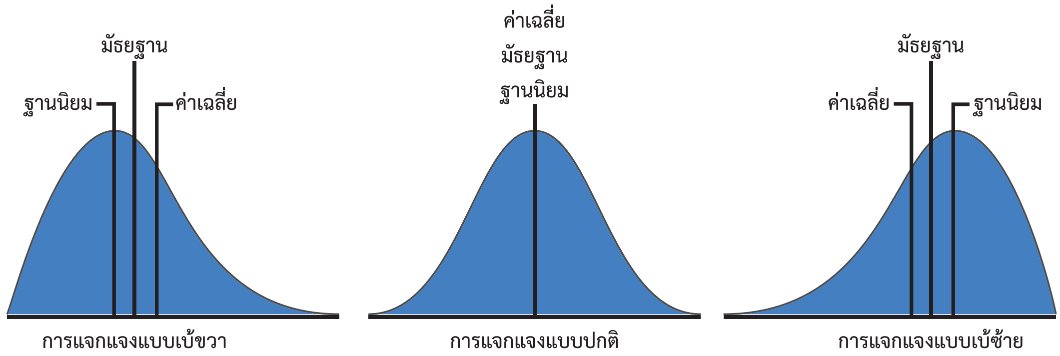
ดังนั้น ค่ามัธยฐานของตัวแปรประสิทธิภาพงาน ($\tilde{x}_{\text{JP}}$) เท่ากับ 11.5

ตัวอย่าง 3.3

พิจารณา ชุดค่าของตัวแปรคะแนนสอบ TOEFL $X_{\text{TOEFL}} = (118, 107, 104, 110, 109, 101, 105, 108, 110, 106, 110, 102, 104, 101, 108)$ จากกลุ่มตัวอย่างผู้เข้าสอบ TOEFL จำนวน 15 คน จะสามารถหาค่ามัธยฐานได้โดยเรียงลำดับข้อมูลจากค่าน้อยไปค่ามาก ได้ดังนี้ $(101, 101, 102, 104, 104, 105, 106, 107, 108, 108, 109, 110, 110, 110, 118)$ ชุดค่าข้อมูลนี้มีจำนวนข้อมูล n เท่ากับ 15 ซึ่งเป็นเลขคี่ ดังนั้น ค่ามัธยฐาน สามารถคำนวณได้โดยใช้สมการ (3.2) ดังนี้

$$\begin{aligned}
\tilde{x}_{\text{TOEFL}} &= x_{\frac{(n+1)}{2},\text{TOEFL}} \\
&= x_{\frac{(15+1)}{2},\text{TOEFL}} \\
&= x_{\frac{(16)}{2},\text{TOEFL}} \\
&= x_{8,\text{TOEFL}} = 107
\end{aligned}$$

ดังนั้น ค่ามัธยฐานของตัวแปรคะแนนสอบ TOEFL ($\tilde{x}_{\text{TOEFL}}$) เท่ากับ 107



ภาพ 3.1: ค่าเฉลี่ย มัธยฐาน และฐานนิยม ของลักษณะการกระจายของข้อมูลที่มีการแจกแจงแบบปกติ การแจกแจงแบบเบ้ขวา และการแจกแจงแบบเบ้ซ้าย

ฐานนิยม (Mode) คือ ค่าข้อมูลที่มีความถี่สูงที่สุด การหาค่าฐานนิยมของตัวแปร X_j แทนด้วยสัญลักษณ์ Mo_j สามารถหาได้โดยการนับความถี่ของค่าข้อมูลแต่ละค่าในชุดของค่าตัวแปร X_j และค่าข้อมูลที่มีความถี่สูงสุดจะเป็นค่าฐานนิยมของค่าตัวแปร X_j ทั้งนี้ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลนามบัญญัติ ข้อมูลเชิงลำดับ ข้อมูลระดับอัตรภาค หรือข้อมูลระดับอัตราส่วน สามารถคำนวณค่าฐานนิยมได้

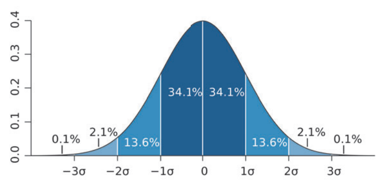
ตัวอย่าง 3.4

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{\text{JP}} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัท จำนวน 10 คน สามารถนับความถี่ของค่าข้อมูล ได้ดังนี้

ค่าข้อมูล	7	**10**	11	12	14	15	16
ความถี่	1	**3**	1	2	1	1	1

จะเห็นว่า ค่าข้อมูล 10 มีความถี่สูงที่สุด นั่นคือ ค่าความถี่เท่ากับ 3 ดังนั้น ค่าฐานนิยมของตัวแปรประสิทธิภาพของงาน (Mo_{JP}) จึงเท่ากับ 10

สำหรับชุดค่าตัวแปรหนึ่งๆ ค่ากลางของข้อมูลทั้ง 3 แบบนั้นอาจจะมีค่าเท่ากันหรือแตกต่างก็ได้ ทั้งนี้ขึ้นอยู่กับลักษณะการแจกแจงของค่าตัวแปรนั้นด้วย ดังแสดงในภาพ 3.1 ค่ากลางของข้อมูลทั้ง 3 แบบจะมีค่าเท่ากันก็ต่อเมื่อค่าข้อมูลมีการแจกแจงแบบปรกติ หากค่าข้อมูลมีลักษณะการแจกแจงแบบเบ้ซ้ายแล้ว ค่าเฉลี่ยจะมีค่าน้อยกว่าค่ามัธยฐาน และค่ามัธยฐานจะมีค่าน้อยกว่าค่าฐานนิยม ในทางกลับกันหากค่าข้อมูลมีลักษณะการแจกแจงแบบเบ้ขวาแล้ว ค่าเฉลี่ยจะมีค่ามากกว่าค่ามัธยฐาน และค่ามัธยฐานจะมีค่ามากกว่าค่าฐานนิยม



ภาพ 3.2: แผนภาพ การ แจกแจง ปรกติ แสดง สัดส่วน พื้นที่ ใต้ โค้ง ที่ ขนาด 1 เท่า 2 เท่า และ 3 เท่า ของ ส่วน เบี่ยง เบน มาตรฐาน ของ ประชากร (σ) ทั้ง ด้าน ซ้าย และ ขวา ของ ค่า เฉลี่ย ประชากร (μ) (ที่มา: https://en.wikipedia.org/wiki/Standard_deviation)

ส่วนเบี่ยงเบนมาตรฐานและความแปรปรวน

ส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) เป็นค่าวัดทางสถิติที่บ่งบอกถึงการกระจายของค่าข้อมูลจากค่าเฉลี่ยไปทางด้านซ้ายหรือขวา ซึ่งเป็นระยะห่างเฉลี่ยกำลังสองของค่าข้อมูลทุกค่าจากค่าเฉลี่ย แสดงตำแหน่งของส่วนเบี่ยงเบนมาตราฐานบนแผนภาพการแจกแจงปรกติ ดังภาพ 3.2 ส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปร X_j ของประชากร แทนด้วยสัญลักษณ์ σ_{X_j} สามารถคำนวณได้โดย

$$\sigma_{X_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \mu_j)^2} \tag{3.4}$$

และส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปร X_j ของตัวอย่าง แทนด้วยสัญลักษณ์ s_{X_j} สามารถหาได้จากสูตร

$$s_{X_j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2} \tag{3.5}$$

ส่วนเบี่ยงเบนมาตรฐานนั้นมีหน่วยเช่นเดียวกับหน่วยของตัวแปร การกระจายของค่าข้อมูลจากค่าเฉลี่ยแปรผันตรงกับค่าส่วนเบี่ยงเบนมาตรฐาน กล่าวคือ ส่วนเบี่ยงเบนมาตรฐานที่มีค่ามากบ่งบอกถึงค่าข้อมูลมีการกระจายตัวมาก ในทางกลับกันส่วนเบี่ยงเบนมาตรฐานมีค่าน้อยแสดงถึงการกระจายของค่าข้อมูลมีน้อยตาม

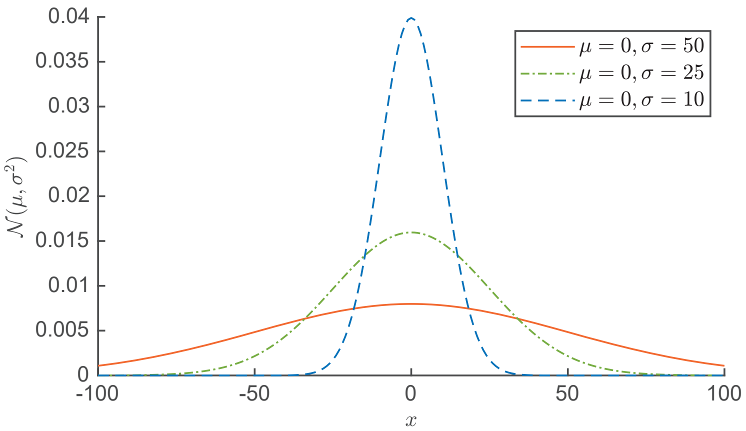
ค่าสถิติที่แสดงลักษณะการกระจายของข้อมูลที่มีความเกี่ยวเนื่องกับส่วนเบี่ยงเบนมาตรฐาน คือ **ความแปรปรวน (Variance)** ซึ่งเป็นค่ากำลังสองของส่วนเบี่ยงเบนมาตรฐาน ความแปรปรวนของค่าตัวแปร X_j ของประชากร แทนด้วยสัญลักษณ์ $\sigma^2_{X_j}$ คำนวณได้โดย

$$\sigma^2_{X_j} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \mu_j)^2 \tag{3.6}$$

และความแปรปรวนของค่าตัวแปร X_j ของตัวอย่าง แทนด้วยสัญลักษณ์ $s^2_{X_j}$ สามารถหาได้จากสูตร

$$s^2_{X_j} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2 \tag{3.7}$$

ค่าความแปรปรวนของข้อมูลมาก แสดงถึงการกระจายของข้อมูลมีมาก ในทางกลับกันค่าความแปรปรวนของข้อมูลน้อย บ่งบอกถึงการกระจายของข้อมูลมีน้อยด้วย ภาพ 3.3 แสดงลักษณะการกระจายของข้อมูลแบบแจกแจงปรกติ 3 ลักษณะ ซึ่งมีค่าส่วนเบี่ยงเบนมาตรฐานและค่าความแปรปรวนต่างกัน ในขณะ



ภาพ 3.3: แผนภาพ การ แจกแจง ปรกติ 3 ลักษณะ ที่ มี ค่า ส่วน เบี่ยง เบน มาตรฐาน (σ) เท่ากับ 50 25 และ 10 และค่าเฉลี่ยประชากร (μ) เท่ากับ 0

ที่ค่าเฉลี่ยเลขคณิตมีค่าเท่ากัน

ตัวอย่าง 3.5

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{\text{JP}} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัท จำนวน 10 คน ซึ่งมีค่าเฉลี่ย ($\overline{x}_{\text{JP}}$) เท่ากับ 11.7 (จากตัวอย่าง 3.1) จะสามารถหาค่าส่วนเบี่ยงเบนมาตรฐานและค่าความแปรปรวน ได้ดังนี้

$x_{i,\text{JP}}$	$x_{i,\text{JP}} - \overline{x}_{\text{JP}}$	$(x_{i,\text{JP}} - \overline{x}_{\text{JP}})^2$
7	7-11.7 = -4.7	22.09
10	10-11.7 = -1.7	2.89
11	11-11.7 = -0.7	0.49
15	15-11.7 = 3.3	10.89
10	10-11.7 = -1.7	2.89
10	10-11.7 = -1.7	2.89
12	12-11.7 = 0.3	0.09
14	14-11.7 = 2.3	5.29
16	16-11.7 = 4.3	18.49
12	12-11.7 = 0.3	0.09
	รวม	66.10

จากสูตรการหาค่าความแปรปรวน

$$s^2_{X_{\text{JP}}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,\text{JP}} - \overline{x}_{\text{JP}})^2$$

แทนค่าในสูตร จะได้

$$s^2_{X_{\mathrm{JP}}} = \frac{1}{10-1} \times 66.10$$
$$= \frac{1}{9} \times 66.10 = 7.34$$

และจากสูตรการหาค่าส่วนเบี่ยงเบนมาตรฐาน

$$s_{X_{\mathrm{JP}}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}})^2}$$

แทนค่าในสูตรจะได้

$$s_{X_{\mathrm{JP}}} = \sqrt{\frac{1}{10-1} \times 66.10}$$
$$= \sqrt{\frac{1}{9} \times 66.10} = \sqrt{7.34} = 2.71$$

ดังนั้น ค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรประสิทธิภาพของงาน ($s_{X_{\mathrm{JP}}}$) เท่ากับ 2.71 และค่าความแปรปรวนของตัวแปรประสิทธิภาพของงาน ($s^2_{X_{\mathrm{JP}}}$) เท่ากับ 7.34 แสดงให้เห็นว่า ค่าข้อมูลประสิทธิภาพของงานส่วนใหญ่กระจายจากค่าเฉลี่ย ±2.71 หน่วย

ความเบ้และความโด่ง

ความเบ้ (Skewness) คือ ความไม่สมมาตรของลักษณะการแจกแจงค่าตัวแปร สามารถแบ่งออกได้ 2 ลักษณะ ได้แก่ การแจกแจงแบบเบ้ขวา (Positive Skewness) และ การแจกแจงแบบเบ้ซ้าย (Negative Skewness) แสดงลักษณะเส้นโค้งการแจกแจงแบบเบ้ขวา และเบ้ซ้าย ดังภาพ 3.1 นอกจากเราจะสามารถระบุความเบ้ของลักษณะการแจกแจงของค่าตัวแปรได้การสังเกตจากค่ากลางข้อมูลแล้ว เรายังสามารถระบุได้จากการวัดค่าสัมประสิทธิ์ความเบ้ การวัดค่าสัมประสิทธิ์ความเบ้นั้นมีวิธีการคำนวณหลายวิธี ในที่นี้จะกล่าวเฉพาะการวัดค่าสัมประสิทธิ์ความเบ้ด้วยวิธีโมเมนต์ Joanes and Gill 1998 ค่าสัมประสิทธิ์ความเบ้ของค่าตัวแปร X_j แทนด้วยสัญลักษณ์ g_1 สามารถคำนวณ โดยใช้สูตรต่อไปนี้

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2\right]^{\frac{3}{2}}} \tag{3.8}$$

เมื่อ $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^3$ คือ โมเมนต์ศูนย์กลาง ที่ 3 ของตัวแปร X_j และ $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2$ คือ โมเมนต์ศูนย์กลาง ที่ 2 ของตัวแปร X_j ถ้าค่า

(Joanes and Gill 1998) D. N. Joanes and C. A. Gill (1998). “Comparing measures of sample skewness and kurtosis.” In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1. url: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9884.00122

สัมประสิทธิ์ความเบ้ เท่ากับ 0 แสดงว่าค่าตัวแปรมีลักษณะการแจกแจงปกติ ถ้าค่าสัมประสิทธิ์ความเบ้ น้อยกว่า 0 แสดงว่าค่าตัวแปรมีลักษณะการแจกแจงแบบเบ้ซ้าย ในขณะที่ถ้าสัมประสิทธิ์ความเบ้มีค่ามากกว่า 0 แสดงว่าค่าตัวแปรมีลักษณะการแจกแจงแบบเบ้ขวา

ตัวอย่าง 3.6

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{\mathrm{JP}} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัท จำนวน 10 คน ซึ่งมีค่าเฉลี่ย ($\overline{x}_{\mathrm{JP}}$) เท่ากับ 11.7 (จากตัวอย่าง 3.1) จะสามารถหาค่าสัมประสิทธิ์ความเบ้ด้วยวิธีเมเมนต์ ได้ดังนี้

$x_{i,\mathrm{JP}}$	$x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}}$	$(x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}})^2$	$(x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}})^3$
7	7-11.7 = -4.7	22.09	-103.82
10	10-11.7 = -1.7	2.89	-4.91
11	11-11.7 = -0.7	0.49	-0.34
15	15-11.7 = 3.3	10.89	35.94
10	10-11.7 = -1.7	2.89	-4.91
10	10-11.7 = -1.7	2.89	-4.91
12	12-11.7 = 0.3	0.09	0.03
14	14-11.7 = 2.3	5.29	12.17
16	16-11.7 = 4.3	18.49	79.51
12	12-11.7 = 0.3	0.09	0.03
	รวม	66.10	8.79

จากสูตรการหาค่าสัมประสิทธิ์ความเบ้

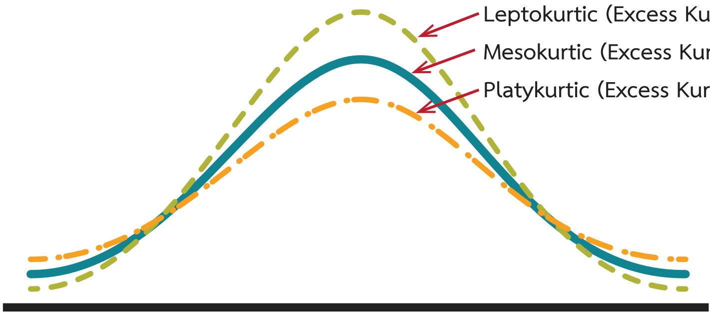
$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)^2\right]^{\frac{3}{2}}}$$

แทนค่าในสูตร จะได้

$$g_1 = \frac{\frac{1}{10} \times 8.79}{\left[\frac{1}{10} \times 66.10\right]^{\frac{3}{2}}}$$

$$= \frac{0.879}{6.610^{\frac{2}{3}}}$$

$$= \frac{0.879}{3.52} = 0.25$$



ภาพ 3.4: แผนภาพการแจกแจง 3 ลักษณะ ตามค่า ค่าอิคเซส-เคอร์โทซิส

ดังนั้น ค่าสัมประสิทธิ์ความเบ้ของตัวแปรประสิทธิภาพของงาน เท่ากับ 0.25 เนื่องจากค่าสัมประสิทธิ์ความเบ้มากกว่า 0 จึงกล่าวได้ว่า ค่าตัวแปรประสิทธิภาพของงานมีลักษณะการแจกแจงแบบเบ้ขวา สอดคล้องกับการพิจารณาค่ากลางข้อมูล ซึ่งมีค่าเฉลี่ยเท่ากับ 11.7 ค่ามัธยฐานเท่ากับ 11.5 และค่าฐานนิยมเท่ากับ 10 จะเห็นว่าค่าเฉลี่ยมีค่ามากกว่าค่ามัธยฐาน และค่ามัธยฐานมีค่ามากกว่าค่าฐานนิยม ซึ่งบ่งบอกถึงการแจกแจงของค่าตัวแปรประสิทธิภาพของงานมีลักษณะเบ้ขวา เช่นกัน

ความโด่ง (Kurtosis) เป็นค่าสถิติที่อธิบายรูปร่างของลักษณะการแจกแจงอีกค่าหนึ่ง ที่แสดงถึงความหนาของปลายหางทั้ง 2 ข้างของโค้งการแจกแจงข้อมูล Westfall 2014 เราสามารถหาค่าความโด่งของตัวแปร X_j แทนด้วยสัญลักษณ์ g_2 ด้วยวิธีโมเมนต์ ได้จากสูตรต่อไปนี้

(Westfall 2014) Peter H. Westfall (2014). “Kurtosis as Peakedness.” In: *The American Statistician* 68.3. url: https://doi.org/10.1080/00031305.2014.917055

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j}-\overline{x}_j)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{i,j}-\overline{x}_j)^2\right]^2} - 3 \tag{3.9}$$

เมื่อ $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j}-\overline{x}_j)^4$ คือ โมเมนต์ศูนย์กลาง ที่ 4 ของตัวแปร X_j และ $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j}-\overline{x}_j)^2$ คือ โมเมนต์ศูนย์กลาง ที่ 2 ของตัวแปร X_j ค่าความโด่งที่คำนวณได้จากสมการ (3.9) เรียกว่า ค่าอิคเซส-เคอร์โทซิส (Excess Kurtosis) ลักษณะความโด่งของโค้งการแจกแจงแบ่งออกได้ 3 ลักษณะ ตามค่าอิคเซส-เคอร์โทซิส (แสดงดังภาพ 3.4) ดังนี้

- พลาติเคอร์ติค (Platykurtic) เป็นการแจกแจงที่มีความโด่งต่ำกว่าความโด่งของการแจกแจงปกติ ซึ่งจะมีค่าอิคเซส-เคอร์โทซิส น้อยกว่า 0
- เมโซเคอร์ติค (Mesokurtic) เป็นการแจกแจงที่มีความโด่งเท่ากับความโด่งของการแจกแจงปกติ ซึ่งจะมีค่าค่าอิคเซส-เคอร์โทซิส เท่ากับ 0
- เลปโตเคอร์ติค (Leptokurtic) เป็นการแจกแจงที่มีความโด่งมากกว่าความโด่งของการแจกแจงปกติ ซึ่งจะมีค่าค่าอิคเซส-เคอร์โทซิส มากกว่า 0

ตัวอย่าง 3.7

พิจารณา ชุดค่าของตัวแปรประสิทธิภาพของงาน (Job Performance: JP) $X_{\mathrm{JP}} = (7, 10, 11, 15, 10, 10, 12, 14, 16, 12)$ จากกลุ่มตัวอย่างพนักงานบริษัท จำนวน 10 คน ซึ่งมีค่าเฉลี่ย ($\bar{x}_{\mathrm{JP}}$) เท่ากับ 11.7 (จากตัวอย่าง 3.1) จะสามารถหาค่าอิคเซส-เคอร์โทซิส ได้ดังนี้

$x_{i,\mathrm{JP}}$	$x_{i,\mathrm{JP}} - \bar{x}_{\mathrm{JP}}$	$(x_{i,\mathrm{JP}} - \bar{x}_{\mathrm{JP}})^2$	$(x_{i,\mathrm{JP}} - \bar{x}_{\mathrm{JP}})^4$
7	7-11.7 = -4.7	22.09	487.97
10	10-11.7 = -1.7	2.89	8.35
11	11-11.7 = -0.7	0.49	0.24
15	15-11.7 = 3.3	10.89	118.59
10	10-11.7 = -1.7	2.89	8.35
10	10-11.7 = -1.7	2.89	8.35
12	12-11.7 = 0.3	0.09	0.01
14	14-11.7 = 2.3	5.29	27.98
16	16-11.7 = 4.3	18.49	341.88
12	12-11.7 = 0.3	0.09	0.01
	รวม	66.10	1001.73

จากสูตรการหาค่าสัมประสิทธิ์ความเบ้

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)^2\right]^2} - 3$$

แทนค่าในสูตร จะได้

$$\begin{aligned} g_2 &= \frac{\frac{1}{10} \times 1001.73}{\left[\frac{1}{10} \times 66.10\right]^2} - 3 \\ &= \frac{100.173}{6.610^2} - 3 \\ &= \frac{100.173}{43.69} - 3 \\ &= 2.29 - 3 = -0.71 \end{aligned}$$

ดังนั้น ค่าอิคเซส-เคอร์โทซิสของตัวแปรประสิทธิภาพของงาน เท่ากับ -0.71 จากค่าอิคเซส-เคอร์โทซิสมีค่าน้อยกว่า 0 จึงกล่าวได้ว่า โค้งการแจกแจงของค่าตัวแปรประสิทธิภาพของงานมีความโด่งน้อยกว่าการแจกแจงปกติ และมีลักษณะความโด่งแบบพลาติเคอร์ติด

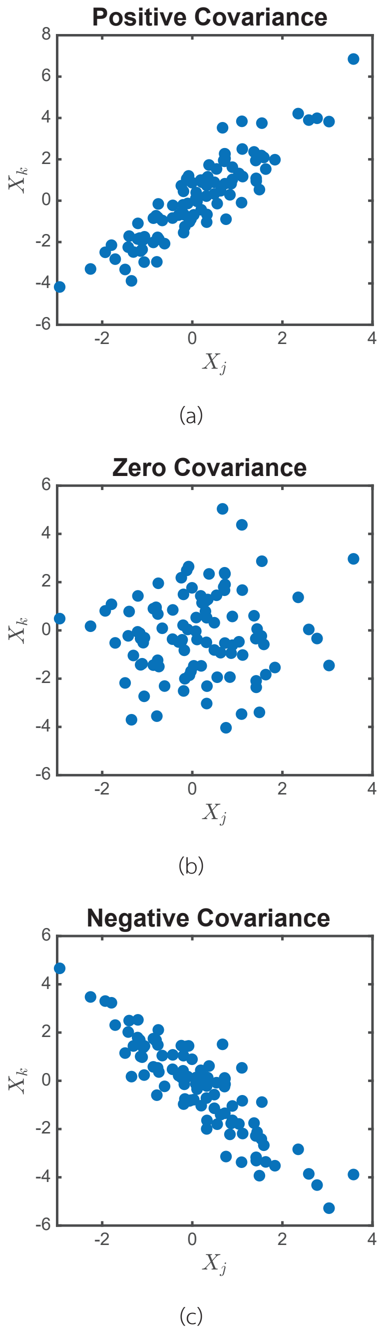
ความแปรปรวนร่วม

ความแปรปรวนร่วม (Covariance) เป็นค่าชี้วัดการเปลี่ยนแปลงของค่าตัวแปรตัวใดตัวหนึ่งเมื่อมีการเปลี่ยนแปลงในค่าตัวแปรอีกตัวหนึ่ง ซึ่งแสดงระดับความสัมพันธ์เชิงเส้นตรงของตัวแปร 2 ตัวแปร กำหนดให้ $X_j = (x_{1,j}, x_{2,j}, ..., x_{n,j})$ และ $X_k = (x_{1,k}, x_{2,k}, ..., x_{n,k})$ เป็นชุดของค่าตัวแปร X_j และ X_k ของข้อมูลจำนวน n ข้อมูล โดยค่า $x_{i,j}$ และ $x_{i,k}$ เป็นค่าตัวแปร X_j และ X_k ของข้อมูลที่ i ค่าความแปรปรวนร่วมระหว่างตัวแปร X_j และ X_k ของประชากร แทนด้วยสัญลักษณ์ σ_{X_j,X_k} สามารถคำนวณได้จากสูตร

$$\sigma_{X_j,X_k} = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \mu_j)(x_{i,k} - \mu_k) \qquad (3.10)$$

และค่าความแปรปรวนร่วมระหว่างตัวแปร X_j และ X_k ของตัวอย่าง แทนด้วยสัญลักษณ์ s_{X_j,X_k} สามารถคำนวณได้จากสูตร

$$s_{X_j,X_k} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - \overline{x}_j)(x_{i,k} - \overline{x}_k) \qquad (3.11)$$



ภาพ 3.5: แผนภาพการกระจายของค่าข้อมูลที่สัมพันธ์กับค่าความแปรปรวนร่วม 3 ลักษณะ คือ (a) ค่าความแปรปรวนร่วมมีค่ามากกว่า 0 (b) ค่าความแปรปรวนร่วมมีค่าเท่ากับ 0 (c) ค่าความแปรปรวนร่วมมีค่าน้อยกว่า 0

ค่าความแปรปรวนร่วม มีค่าที่เป็นไปได้อยู่ระหว่าง $-\infty$ และ ∞ ค่าความแปรปรวนร่วมที่มีค่ามากกว่า 0 แสดงถึงความสัมพันธ์ระหว่างตัวแปรทั้งสอง โดยเมื่อค่าของตัวแปรหนึ่งมีค่ามาก ค่าของอีกตัวแปรหนึ่งมักมีค่ามากด้วย ในทางเดียวกันเมื่อค่าตัวแปรหนึ่งมีค่าน้อย ค่าของอีกตัวแปรหนึ่งมักมีค่าน้อยตามไปด้วย ในกรณีที่ค่าความแปรปรวนร่วมมีค่าน้อยกว่า 0 แสดงถึงความสัมพันธ์ระหว่างสองตัวแปรในทางตรงกันข้าม กล่าวคือเมื่อค่าของตัวแปรหนึ่งมีค่ามาก ค่าของอีกตัวแปรหนึ่งมักมีค่าน้อย ในทางกลับกันเมื่อค่าตัวแปรหนึ่งมีค่าน้อย ค่าของอีกตัวแปรหนึ่งมักมีค่ามาก ส่วนค่าความแปรปรวนร่วมที่มีค่าเท่ากับ 0 บ่งบอกถึงค่าของตัวแปรทั้งสองตัวแปรไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน แสดงแผนภาพการกระจายของค่าข้อมูลระหว่าง 2 ตัวแปร และค่าความแปรปรวนร่วมที่สัมพันธ์กัน ดังภาพ 3.5

เนื่องจากค่าความแปรปรวนร่วมสามารถแสดงระดับความสัมพันธ์ระหว่างตัวแปรได้เพียง 2 ตัวแปร เราจึงไม่สามารถหาค่าความแปรปรวนร่วมเพียงค่าเดียว ในกรณีที่จำนวนตัวแปรที่สนใจมีมากกว่า 2 ตัวแปรได้ อย่างไรก็ตาม เราสามารถพิจารณาค่าความแปรปรวนร่วมของแต่ละคู่ของตัวแปรที่ทำการศึกษา โดยสร้างเมทริกซ์ความแปรปรวนร่วม (Covariance Matrix) ขนาด $d \times d$ เมื่อ d คือ จำนวนตัวแปรที่ศึกษา กำหนด $\mathbf{X} = (X_1, X_2, ..., X_d)$ เป็นชุดตัวแปรบนชุดข้อมูล $\mathbf{D}$ เมทริกซ์ความแปรปรวนร่วมของชุดตัวแปร $\mathbf{X}$ แทนด้วยสัญลักษณ์

$\mathrm{cov}(\mathbf{X})$ สามารถสร้างได้ ดังนี้

$$\mathrm{cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{X_1,X_1} & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_d} \\ \sigma_{X_2,X_1} & \sigma_{X_2,X_2} & \cdots & \sigma_{X_2,X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_d,X_1} & \sigma_{X_d,X_2} & \cdots & \sigma_{X_d,X_d} \end{bmatrix} \tag{3.12}$$

จากคุณสมบัติของค่าความแปรปรวนร่วม $\sigma_{X_j,X_k} = \sigma_{X_k,X_j}$ จึงทำให้เมทริกซ์ความแปรปรวนร่วมเป็นเมทริกซ์สมมาตร (Symmetric Matrix) ซึ่งมีค่าในตำแหน่ง (j,k) เท่ากับค่าในตำแหน่ง (k,j) และจากคุณสมบัติ $\sigma_{X_j,X_j} = \sigma^2_{X_j}$ ทำให้ค่าในเส้นทแยงมุม (ตำแหน่ง (j,k) ที่ $j = k$) ของเมทริกซ์ แสดงค่าความแปรปรวนของค่าตัวแปร X_j นั่นเอง

ตัวอย่าง 3.8

พิจารณา ชุดข้อมูลจากการสำรวจระดับเชาว์ปัญญา (IQ) และประสิทธิภาพของงาน (Job Performance: JP) จากตัวอย่างพนักงานบริษัท จำนวน 10 คน (แสดงดังตาราง ด้านล่าง) โดยค่าเฉลี่ยระดับเชาว์ปัญญา ($\overline{x}_{\mathrm{IQ}}$) เท่ากับ 111.5 และค่าเฉลี่ยประสิทธิภาพของงาน ($\overline{x}_{\mathrm{JP}}$) เท่ากับ 11.7 จะสามารถหาค่าความแปรปรวนร่วมระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน ได้ดังนี้

$x_{i,\mathrm{IQ}}$	$x_{i,\mathrm{JP}}$	$x_{i,\mathrm{IQ}} - \overline{x}_{\mathrm{IQ}}$	$x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}}$	$(x_{i,\mathrm{IQ}} - \overline{x}_{\mathrm{IQ}})(x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}})$
99	7	-12.5	-4.7	58.75
105	10	-6.5	-1.7	11.05
105	11	-6.5	-0.7	4.55
106	15	-5.5	3.3	18.15
108	10	-3.5	-1.7	5.95
112	10	0.5	-1.7	-0.85
113	12	1.5	0.3	0.45
115	14	3.5	2.3	8.05
118	16	6.5	4.3	27.95
134	12	22.5	0.3	6.75
			รวม	140.80

จากสูตรการหาค่าความแปรปรวนร่วมของตัวอย่าง

$$s_{X_{\mathrm{IQ}},X_{\mathrm{JP}}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,\mathrm{IQ}} - \overline{x}_{\mathrm{IQ}})(x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}})$$

แทนค่าในสูตร จะได้

$$
\begin{aligned}
s_{X_{\mathrm{IQ}},X_{\mathrm{JP}}} &= \frac{1}{10-1}\sum_{i=1}^{10}(x_{i,\mathrm{IQ}} - \overline{x}_{\mathrm{IQ}})(x_{i,\mathrm{JP}} - \overline{x}_{\mathrm{JP}}) \\
&= \frac{1}{9} \times 140.80 \\
&= 15.64
\end{aligned}
$$

ดังนั้น ค่าความแปรปรวนร่วมระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน เท่ากับ 15.64 บ่งชี้ว่าค่าตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงานมีความสัมพันธ์กัน โดยเมื่อค่าตัวแปรระดับเชาว์ปัญญามีค่ามาก ค่าตัวแปรประสิทธิภาพของงานจะมีค่ามากด้วย ในขณะเดียวกันเมื่อค่าตัวแปรระดับเชาว์ปัญญามีค่าน้อย ค่าตัวแปรประสิทธิภาพของงานจะมีค่าน้อยด้วย

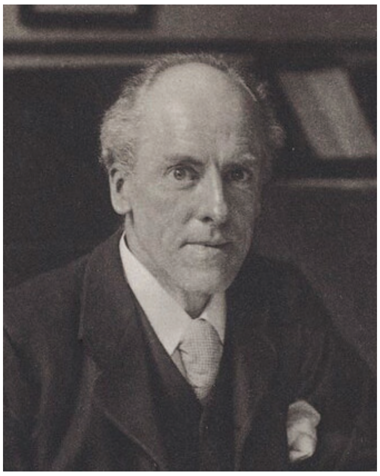
อย่างไรก็ตามขนาดของค่าความแปรปรวนร่วมนั้นยากที่นำมาแปลความหมายและเปรียบเทียบกัน เนื่องจากค่าความแปรปรวนร่วมไม่ได้ถูกทำให้เป็นบรรทัดฐาน อีกทั้งยังขึ้นอยู่กับขนาดของค่าตัวแปร ค่าวัดระดับความสัมพันธ์เชิงเส้นตรงของตัวแปร 2 ตัวแปร อีกค่าหนึ่ง คือ **ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Coefficient)** ซึ่งเป็นค่าความแปรปรวนร่วมระหว่างตัวแปร 2 ตัวแปรที่ถูกทำให้เป็นบรรทัดฐาน โดยการหารด้วยผลคูณของค่าส่วนเบี่ยงเบนมาตรฐานของทั้งสองตัวแปร ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปร X_j และ X_k สำหรับประชากร แทนด้วยสัญลักษณ์ ρ_{X_j,X_k} สามารถคำนวณได้จาก

$$\rho_{X_j,X_k} = \frac{\sigma_{X_j,X_k}}{\sigma_{X_j}\sigma_{X_k}} \qquad (3.13)$$

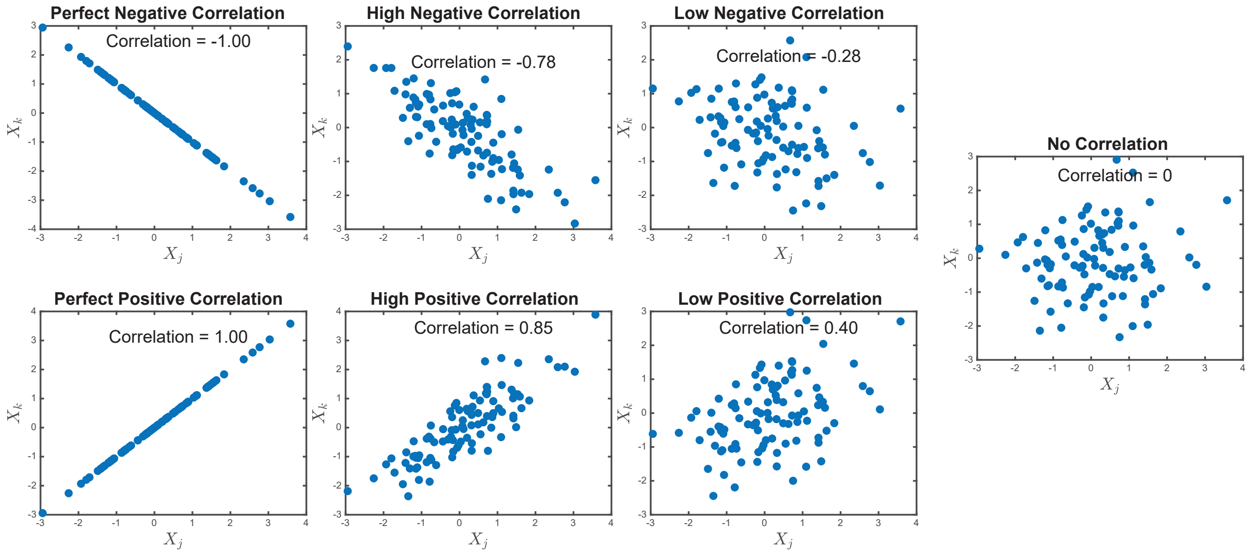
และ ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปร X_j และ X_k สำหรับตัวอย่าง แทนด้วยสัญลักษณ์ r_{X_j,X_k} สามารถคำนวณได้จาก

$$r_{X_j,X_k} = \frac{s_{X_j,X_k}}{s_{X_j}s_{X_k}} \qquad (3.14)$$

ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน มีค่าที่เป็นไปได้อยู่ระหว่าง -1 และ 1 ค่าสัมประสิทธิ์สหสัมพันธ์ ที่มีค่าเท่ากับ 1 บ่งบอกถึงความสัมพันธ์เชิงเส้นตรงอย่างสมบูรณ์แบบระหว่างค่าตัวแปรทั้งสอง เมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้นค่าของอีกตัวแปรหนึ่งจะเพิ่มขึ้นตามไปด้วย ส่วนค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่าเท่ากับ -1 แสดงถึงความสัมพันธ์เชิงเส้นตรงในลักษณะตรงข้ามอย่างสมบูรณ์ของตัวแปรทั้งสอง โดยเมื่อค่าของตัวแปรหนึ่งเพิ่มขึ้นค่าของอีกตัวแปรหนึ่งจะลดลง ในกรณีที่ค่าสัมประสิทธิ์สหสัมพันธ์มีค่าเท่ากับ 0 แสดงถึงค่าของตัวแปรทั้งสองตัวแปรไม่มีความสัมพันธ์เชิงเส้นตรงต่อกัน ค่าสัมบูรณ์ (Abso-



ภาพ 3.6: คาร์ล เพียร์สัน (Karl Pearson: ค.ศ. 1857-1936) นักคณิตศาสตร์และนักชีวสถิติชาวอังกฤษ ผู้พัฒนาสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (ที่มา: https://en.wikipedia.org/wiki/Karl_Pearson)



ภาพ 3.7: แผนภาพการกระจายของค่าข้อมูลที่สัมพันธ์กับค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน 7 ลักษณะ

lute Value) ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ยังชี้วัดระดับความสัมพันธ์ของค่าตัวแปรทั้งสองด้วย กล่าวคือ ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่ามาก บ่งบอกถึงค่าตัวแปรทั้งสองมีความสัมพันธ์กันอย่างมากกว่า ในขณะที่ค่าสัมบูรณ์ของค่าสัมประสิทธิ์สหสัมพันธ์ที่มีค่าน้อย แสดงถึงความสัมพันธ์ในระดับต่ำระหว่างตัวแปรทั้งสองด้วย แสดงแผนภาพการกระจายของค่าข้อมูลระหว่าง 2 ตัวแปร และสัมประสิทธิ์สหสัมพันธ์เพียร์สันที่สัมพันธ์กัน ดังภาพ 3.7

เช่น เดียว กับ ค่า ความ แปรปรวน ร่วม เรา สามารถ อธิบาย ความ สัมพันธ์สำหรับชุดข้อมูลที่มีจำนวนตัวแปรมากกว่า 2 ตัวแปรได้ โดยการสร้างเมทริกซ์สหสัมพันธ์เพียร์สัน (Pearson Correlation Matrix) สำหรับ อธิบาย ความสัมพันธ์ระหว่างแต่ละคู่ของตัวแปร เมทริกซ์สหสัมพันธ์เพียร์สัน ของชุดตัวแปร X แทนด้วยสัญลักษณ์ corr(X) สามารถสร้างได้ ดังนี้

$$\mathrm{corr}(\mathbf{X}) = \begin{bmatrix} \rho_{X_1,X_1} & \rho_{X_1,X_2} & \cdots & \rho_{X_1,X_d} \\ \rho_{X_2,X_1} & \rho_{X_2,X_2} & \cdots & \rho_{X_2,X_d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{X_d,X_1} & \rho_{X_d,X_2} & \cdots & \rho_{X_d,X_d} \end{bmatrix} \qquad (3.15)$$

เมทริกซ์สหสัมพันธ์เพียร์สันเป็นเมทริกซ์สมมาตร (Symmetric Matrix) ซึ่งมีค่าในตำแหน่ง (j,k) เท่ากับค่าในตำแหน่ง (k,j) และค่าในเส้นทแยงมุมของเมทริกซ์ ซึ่งแสดงค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันภายในตัวแปรเดียวกัน มีค่าเท่ากับ 1 เสมอ

ตัวอย่าง 3.9

พิจารณา ชุดข้อมูลจากการสำรวจระดับเชาว์ปัญญา (IQ) และประสิทธิภาพของงาน (Job Performance: JP) จากตัวอย่างพนักงานบริษัท จำนวน 10 คน โดยค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปรระดับเชาว์ปัญญา (s_{IQ}) เท่ากับ 9.70 ค่าส่วนเบี่ยงเบนมาตรฐานของค่าตัวแปรประสิทธิภาพของงาน (s_{JP}) เท่ากับ 2.71 และค่าความแปรปรวนร่วมระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน เท่ากับ 15.64 (จากตัวอย่าง 3.8) จะสามารถหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน ได้ดังนี้

จากสูตรการหาค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันของตัวอย่าง

$$r_{X_{\mathrm{IQ}},X_{\mathrm{JP}}} = \frac{s_{\mathrm{IQ},X_{\mathrm{JP}}}}{s_{\mathrm{IQ}}s_{\mathrm{JP}}} \tag{3.16}$$

แทนค่าในสูตร จะได้

$$\begin{aligned} r_{X_{\mathrm{IQ}},X_{\mathrm{JP}}} &= \frac{15.64}{9.70 \times 2.71} \\ &= \frac{15.64}{26.287} \\ &= 0.59 \end{aligned}$$

ดังนั้น ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรระดับเชาว์ปัญญาและประสิทธิภาพของงาน เท่ากับ 0.59 บ่งชี้ว่าค่าตัวแปรระดับเชาว์ปัญญา และประสิทธิภาพของงานมีความสัมพันธ์เชิงเส้นตรงต่อกัน โดยเมื่อค่าตัวแปรระดับเชาว์ปัญญามีค่าเพิ่มขึ้น ค่าตัวแปรประสิทธิภาพของงานจะมีค่าเพิ่มขึ้นตามด้วย ในขณะเดียวกันเมื่อค่าตัวแปรระดับเชาว์ปัญญามีค่าลดลง ค่าตัวแปรประสิทธิภาพของงานจะมีค่าลดลงด้วย

ทั้งค่าความแปรปรวนร่วมและค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน เป็นการหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร สามารถนำไปใช้เป็นเครื่องมือในการวิเคราะห์เชิงวินิจฉัยได้ อย่างไรก็ตามวิธีการวิเคราะห์ความสัมพันธ์ทั้ง 2 วิธี จะสามารถใช้ได้ก็ต่อเมื่อตัวแปรทั้ง 2 ตัวแปรมีชนิดข้อมูลเป็นข้อมูลเชิงตัวเลขเท่านั้น ยังมีวิธีการทางสถิติอื่นๆ ที่สามารถใช้ในการวัดระดับความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร ให้สามารถเลือกใช้ได้อย่างเหมาะสม เช่น การวิเคราะห์ความแปรปรวนหรืออะโนวา (Analysis of Variance: ANOVA) ซึ่งสามารถใช้ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลขและตัวแปรเชิงกลุ่มได้ และการทดสอบไคกำลังสอง (Chi-squared Test) ซึ่งสามารถใช้ทดสอบความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปรที่มีชนิดข้อมูลเป็นข้อมูลเชิงกลุ่มได้ เป็นต้น

3.2 การวิเคราะห์จัดกลุ่ม

การวิเคราะห์จัดกลุ่ม (Cluster Analysis) เป็นการแบ่งข้อมูลออกเป็นกลุ่ม (Cluster) โดยพิจารณาจากค่าตัวแปรหรือคุณลักษณะที่สนใจ อาจแบ่งกลุ่มข้อมูลจากการพิจารณาเพียงตัวแปรเดียวหรือหลายตัวแปรก็ได้ โดยข้อมูลในกลุ่มเดียวกันจะมีคุณลักษณะคล้ายกัน ส่วนข้อมูลที่อยู่ต่างกลุ่มกันจะมีคุณลักษณะแตกต่างกัน ในบริบทของวิทยาการข้อมูล การจัดกลุ่มข้อมูลถูกดำเนินการอย่างอัตโนมัติโดยใช้วิธีการทางคณิตศาสตร์หรือสถิติ ทำให้สามารถวิเคราะห์จัดกลุ่มของข้อมูลจำนวนมาก พร้อมกับการพิจารณาคุณลักษณะหลายคุณลักษณะร่วมกันได้อย่างมีประสิทธิภาพ ในที่นี้จะกล่าวถึงวิธีการวิเคราะห์แบ่งกลุ่มพื้นฐานและเป็นที่รู้จักอย่างแพร่หลาย 3 วิธีการ ที่มีแนวคิดพื้นฐานที่แตกต่างกัน อย่างไรก็ตามเนื่องจากวิธีการวิเคราะห์จัดกลุ่มด้วยวิธีการทางคณิตศาสตร์หรือสถิติมีพื้นฐานบนการวัดระยะห่างระหว่างข้อมูล 2 ข้อมูล จึงจะได้กล่าวถึงการวัดระยะห่างระหว่างข้อมูลก่อนกล่าวถึงวิธีการวิเคราะห์กลุ่มต่อไป ทั้งนี้พึงต้องทราบไว้ประการหนึ่งว่า วิธีการวิเคราะห์จัดกลุ่ม โครงแบบของวิธีการวิเคราะห์แบ่งกลุ่ม และตัวแปรที่นำมาพิจารณา ที่แตกต่างกัน อาจทำให้ได้ผลลัพธ์การแบ่งกลุ่มที่แตกต่างกันด้วย อีกทั้งอาจไม่สอดคล้องกับกลุ่มข้อมูลที่ถูกจัดโดยมนุษย์ด้วยเช่นกัน

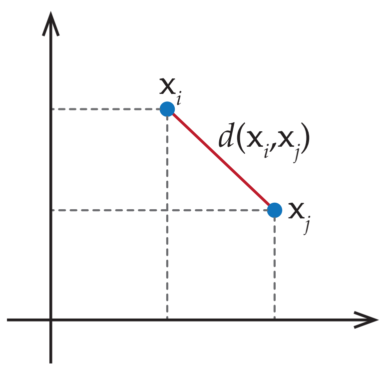
การวัดระยะห่างระหว่างข้อมูล

กำหนดให้ $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$ และ $\mathbf{x}_j = (x_{j,1}, x_{j,2}, ..., x_{j,d})$ เป็นข้อมูล 2 ข้อมูลที่ประกอบไปด้วยตัวแปรจำนวน d ตัวแปร เราสามารถแสดงระยะห่างระหว่างข้อมูล 2 ข้อมูล บนปริภูมิลักษณะเด่น d มิติได้ ด้วยค่าระยะทาง (Distance) ระหว่างข้อมูล ซึ่งบ่งบอกถึงขนาดของความแตกต่างระหว่างค่าตัวแปรของข้อมูลทั้ง 2 ข้อมูล โดยทั่วไปแล้ว ค่าระยะทางระหว่างข้อมูลที่มีค่าน้อยแสดงว่าข้อมูลอยู่ใกล้กันบนปริภูมิลักษณะเด่น ในขณะที่ค่าระยะทางที่มีค่ามากบ่งบอกว่าข้อมูลอยู่ไกลกันมากด้วย การวัดระยะทางระหว่างข้อมูล 2 ข้อมูลสามารถทำได้หลายวิธี มีวิธีการที่เป็นที่รู้จักโดยทั่วไป ดังนี้

ระยะทางแบบยุคลิด (Euclidean Distance)

ระยะทางแบบยุคลิด ระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ คือ ความยาวของเส้นตรงที่ลากระหว่างจุดข้อมูลทั้ง 2 จุดข้อมูลในปริภูมิยุคลิด d มิติ สามารถคำนวณได้ ดังนี้

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2} \tag{3.17}$$



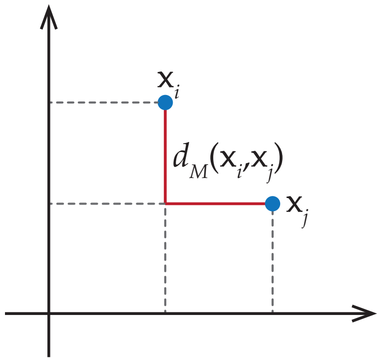
ภาพ 3.8: แนวคิดของการวัดระยะทางแบบยุคลิดของข้อมูลในปริภูมิ 2 มิติ

เมื่อ $d(\mathbf{x}_i, \mathbf{x}_j)$ คือ ค่าระยะทางแบบยุคลิดระหว่างข้อมูลทั้งสองข้อมูล ซึ่งจะมีค่าเท่ากับ 0 เมื่อข้อมูลทั้งสองมีค่าตัวแปรทุกตัวแปรเท่ากัน นั่นหมายถึง ข้อมูลทั้งสองเป็นจุดข้อมูลเดียวกันบนปริภูมิลักษณะเด่น ระยะทางแบบยุคลิดจะมีค่ามาก เมื่อค่าตัวแปรของข้อมูลทั้ง 2 ข้อมูลแตกต่างกันมาก แสดงแนวคิดของการวัดระยะทางแบบยุคลิดของข้อมูลในปริภูมิ 2 มิติ ดังภาพ 3.8

ระยะทางแบบแมนฮัตตัน (Manhattan Distance)

ระยะทางแบบแมนฮัตตัน ระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ แทนด้วย สัญลักษณ์ $d_M(\mathbf{x}_i, \mathbf{x}_j)$ สามารถคำนวณ ได้ดังนี้

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{d} |x_{i,k} - x_{j,k}| \qquad (3.18)$$



ภาพ 3.9: แนวคิดของการวัดระยะทางแบบแมนฮัตตันของข้อมูลในปริภูมิ 2 มิติ

เช่นเดียวกันระยะทางแบบยุคลิด ค่าระยะทางแบบแมนฮัตตัน จะมีค่าเท่ากับ 0 เมื่อข้อมูลทั้ง 2 ข้อมูล มีค่าตัวแปรเท่ากันทุกตัวแปร ระยะทางแบบแมนฮัตตันมีค่ามาก แสดงว่าข้อมูลทั้ง 2 ข้อมูลอยู่ห่างกันมากเช่นกัน แสดงแนวคิดของการวัดระยะทางแบบแมนฮัตตันของข้อมูลในปริภูมิ 2 มิติ ดังภาพ 3.8 ค่าระยะทางแบบแมนฮัตตันนี้เหมือนระยะการเดินทางจากสี่แยกหนึ่งไปยังอีกสี่แยกหนึ่งในเขตการปกครองท้องถิ่นแมนฮัตตัน นครนิวยอร์ก สหรัฐอเมริกา ซึ่งมีการวางผังเมืองเป็นรูปตารางพิกัดฉาก

ภาพ 3.10: แผนที่ ใจกลาง เขตการ ปกครอง ท้อง ถิ่น แมน ฮัต ตัน นคร นิวยอร์ก สหรัฐอเมริกา (ที่มา: https://www.pinterest.com/pin/107523509833448381/)

ตัวอย่าง 3.10

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 4 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (5.1, 3.5, 1.4, 0.2) \text{ และ } \mathbf{x}_2 = (5.0, 3.5, 1.6, 0.6)$$

จะสามารถวัดระยะทางแบบยุคลิดระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ (3.17) ดังนี้

$$\begin{aligned} d(\mathbf{x}_1, \mathbf{x}_2) &= \sqrt{\sum_{k=1}^{4} (x_{1,k} - x_{2,k})^2} \\ &= \sqrt{(5.1 - 5.0)^2 + (3.5 - 3.5)^2 + (1.4 - 1.6)^2 + (0.2 - 0.6)^2} \\ &= \sqrt{0.1^2 + 0^2 + (-0.2)^2 + (-0.4)^2} \\ &= \sqrt{0.01 + 0 + 0.04 + 0.16} \\ &= \sqrt{0.21} = 0.46 \end{aligned}$$

และจะสามารถวัดระยะทางแบบแมนฮัตตันระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดย

ใช้สมการ (3.18) ดังนี้

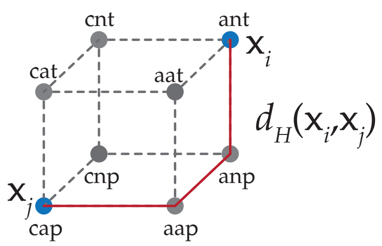
$$d_M(\mathbf{x}_1, \mathbf{x}_2) = \sum_{k=1}^{4} |x_{1,k} - x_{2,k}|$$
$$= |5.1 - 5.0| + |3.5 - 3.5| + |1.4 - 1.6| + |0.2 - 0.6|$$
$$= 0.1 + 0 + 0.2 + 0.4$$
$$= 0.7$$

ดังนั้น ระยะทางแบบยุคลิดและระยะทางแบบแมนฮัตตันระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.46 และ 0.7 ตามลำดับ

ระยะทางแบบแฮมมิง (Hamming Distance)

ระยะทางแบบแฮมมิง เป็นการนับจำนวนตัวแปรที่มีค่าไม่เหมือนกันของข้อมูลทั้ง 2 ข้อมูล ซึ่งมักใช้กับข้อมูลที่มีตัวแปรเป็นชนิดข้อมูลเชิงกลุ่ม สามารถคำนวณได้ ดังนี้

$$d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{d} (x_{i,k} \neq x_{j,k}) \tag{3.19}$$



ภาพ 3.11: แนวคิดของการวัดระยะทางแบบแฮมมิง แสดงการเปลี่ยนคำว่า “cap” ไปเป็นคำว่า “ant” โดยมีลำดับการเปลี่ยนแปลงจากตัวอักษรทางซ้ายไปยังตัวอักษรทางขวา

เมื่อ $d_H(\mathbf{x}_i, \mathbf{x}_j)$ คือ ค่าระยะทางแบบแฮมมิง และการดำเนินการเปรียบเทียบไม่เท่ากับ $\neq$ ให้ผลลัพธ์เป็น 1 ก็ต่อเมื่อ ตัวถูกดำเนินการทั้งสองมีค่าต่างกัน และให้ผลลัพธ์เป็น 0 ก็ต่อเมื่อตัวถูกดำเนินการทั้งสองมีค่าไม่ต่างกัน ค่าระยะทางแบบแฮมมิง จะมีค่าเท่ากับ 0 ก็ต่อเมื่อข้อมูลทั้ง 2 ข้อมูลมีค่าตัวแปรเหมือนกันทุกตัวแปร และมีค่ามากที่สุดที่เป็นไปได้ เท่ากับ จำนวนตัวแปรของข้อมูล ซึ่งจะเกิดขึ้นในกรณีที่ข้อมูลทั้ง 2 ข้อมูลมีค่าตัวแปรแตกต่างกันทุกตัวแปร แนวคิดของการวัดระยะทางแบบแฮมมิง (ซึ่งแสดงตัวอย่างสำหรับคำที่ประกอบด้วย 3 ตัวอักษร ดังภาพ 3.11) สามารถอธิบายได้ด้วยจำนวนครั้งในการแก้ไขคำๆ หนึ่งให้เป็นคำอีกคำหนึ่งที่มีความยาวของคำเท่ากัน (ในภาพ 3.11 แสดงการแก้ไขคำว่า “cap” ไปเป็นคำว่า “ant” จากลำดับตัวอักษรซ้ายสุดไปยังตัวอักษรขวาสุด)

ตัวอย่าง 3.11

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงกลุ่ม 3 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (\text{Short}, \text{Medium}, \text{versicolor}) \text{ และ}$$
$$\mathbf{x}_2 = (\text{VeryShort}, \text{Medium}, \text{setosa})$$

จะสามารถวัดระยะทางแบบแฮมมิงระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ

3.19 ดังนี้

$$d_H(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{d}(x_{i,k} \neq x_{j,k})$$
$$= (\text{Short} \neq \text{VeryShort}) + (\text{Medium} \neq \text{Medium})$$
$$+ (\text{versicolor} \neq \text{setosa})$$
$$= 1 + 0 + 1 = 2$$

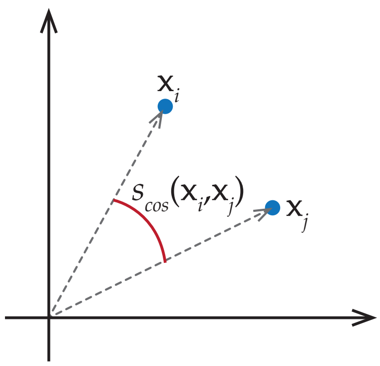
ดังนั้น ระยะทางแบบแฮมมิงระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 2 นั่นคือ มีตัวแปรที่ข้อมูล $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าต่างกัน จำนวน 2 ตัวแปร

นอกจาก ค่า ระยะ ทาง แล้ว เรา ยัง สามารถ การ อธิบาย ระยะ ห่าง ระหว่าง ข้อมูลด้วยค่าความคล้าย (Similarity) ซึ่งบ่งบองขนาดของความใกล้เคียงกัน ระหว่างค่าตัวแปรของข้อมูลทั้งสอง มีวิธีการวัดความคล้ายที่น่าสนใจ ดังนี้

ความคล้ายแบบโคไซน์ (Cosine Similarity)

ความคล้ายแบบโคไซน์ของข้อมูล 2 ข้อมูล เป็นการวัดค่าโคไซน์ (Cosine) ของมุมระหว่างเวกเตอร์ 2 เวกเตอร์ ซึ่งพุ่งจากจุดกำเนิด (Origin) ไปยังจุดข้อมูลทั้งสองบนปริภูมิลักษณะเด่น แนวคิดของการวัดความคล้ายแบบโคไซน์ของข้อมูลในปริภูมิ 2 มิติ สามารถแสดงได้ดังภาพ 3.12 ความคล้ายแบบโคไซน์ระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ แทนด้วยสัญลักษณ์ $s_{cos}(\mathbf{x}_i, \mathbf{x}_j)$ สามารถคำนวณได้ ดังนี้

$$s_{cos}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^{d} x_{i,k}x_{j,k}}{\sqrt{\sum_{k=1}^{d} x_{i,k}^2}\sqrt{\sum_{k=1}^{d} x_{j,k}^2}} \tag{3.20}$$



ภาพ 3.12: แนวคิดของการวัดความคล้ายแบบโคไซน์ของข้อมูลในปริภูมิ 2 มิติ

ความคล้ายแบบโคไซน์จะมีค่าเท่ากับ 1 เมื่อเวกเตอร์ของข้อมูล 2 ข้อมูลมีทิศทางเดียวกัน ความคล้ายแบบโคไซน์จะมีค่าเท่ากับ 0 เมื่อเวกเตอร์ของข้อมูล 2 ข้อมูลทำมุมตั้งฉากกัน และความคล้ายแบบโคไซน์จะมีค่าเท่ากับ -1 เมื่อเวกเตอร์ของข้อมูล 2 ข้อมูลมีทิศทางตรงข้ามกัน ในทางปฏิบัติแล้ว เราจะใช้การวัดความคล้ายแบบโคไซน์ สำหรับข้อมูลที่ค่าตัวแปรทุกตัวมีค่ามากกว่าหรือเท่ากับ 0 ซึ่งจะทำให้ค่าความคล้ายที่วัดได้ อยู่ในช่วงระหว่าง 0 ถึง 1

ตัวอย่าง 3.12

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 4 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (5.1, 3.5, 1.4, 0.2) \text{ และ } \mathbf{x}_2 = (5.0, 3.5, 1.6, 0.6)$$

จะสามารถวัดความคล้ายแบบโคไซน์ระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ 3.20 ดังนี้

$$
\begin{aligned}
s_{cos}(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\sum_{k=1}^{4} x_{1,k}x_{2,k}}{\sqrt{\sum_{k=1}^{4} x_{1,k}^2}\sqrt{\sum_{k=1}^{4} x_{2,k}^2}} \\
&= \frac{(5.1 \times 5.0) + (3.5 \times 3.5) + (1.4 \times 1.6) + (0.2 \times 0.6)}{\sqrt{5.1^2 + 3.5^2 + 1.4^2 + 0.2^2}\sqrt{5.0^2 + 3.5^2 + 1.6^2 + 0.6^2}} \\
&= \frac{25.50 + 12.25 + 2.24 + 0.12}{\sqrt{26.01 + 12.25 + 1.96 + 0.04}\sqrt{25.00 + 12.25 + 2.56 + 0.36}} \\
&= \frac{40.11}{\sqrt{40.26}\sqrt{40.17}} \\
&= \frac{40.11}{6.35 \times 6.34} \\
&= \frac{40.11}{40.26} = 0.9963
\end{aligned}
$$

ดังนั้น ความคล้ายแบบโคไซน์ระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.9963

สัมประสิทธิ์ความคล้ายแจ็คการ์ด (Jaccard Similarity Coefficient)

สัมประสิทธิ์ความคล้ายแจ็คการ์ด เป็นมาตรวัดความคล้ายระหว่างเซตจำกัด (Finite Set) 2 เซต สามารถคำนวณได้ ดังนี้

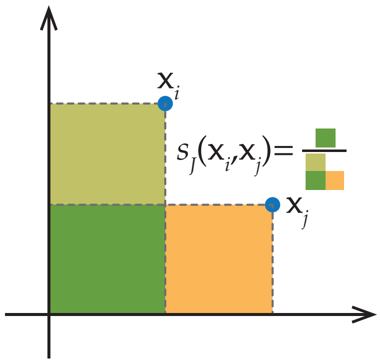
$$s_J(A, B) = \frac{A \cap B}{A \cup B} \tag{3.21}$$

เมื่อ $s_J(A, B)$ คือ ค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดระหว่างเซต A และ B จะมีค่าอยู่ระหว่าง 0 และ 1 โดยถ้าเซตทั้งสองมีสมาชิกร่วมกันจำนวนมากเมื่อเทียบกับจำนวนรวมของสมาชิกของทั้ง 2 เซต ค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดจะมีค่าเข้าใกล้ค่า 1 ในกรณีที่เซตทั้งสองเป็เซตว่าง (Empty Set) ค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดถูกกำหนดให้มีค่าเท่ากับ 1 และหากเซตทั้งสองไม่มีสมาชิกร่วมกันเลย ค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดจะมีค่าเท่ากับ 0

สำหรับข้อมูล 2 ข้อมูล ที่ประกอบด้วยตัวแปรเชิงกลุ่ม เราสามารถหาค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดได้ โดยการเข้ารหัสข้อมูลด้วยวิธีการเข้ารหัสแบบวัน-ฮอทก่อน ซึ่งจะทำให้ข้อมูลมีลักษณะเป็นข้อมูลแบบทวิภาค แล้วจึงคำนวณค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ด ด้วยสูตรต่อไปนี้

$$s_J(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{k=1}^{d} \min(x_{i,k}, x_{j,k})}{\sum_{k=1}^{d} \max(x_{i,k}, x_{j,k})} \tag{3.22}$$

เมื่อ $\min(a, b)$ และ $\max(a, b)$ เป็นฟังก์ชันการหาค่าน้อยที่สุดและค่ามากที่สุด



ภาพ 3.13: แนวคิดของการวัดความคล้ายด้วยสัมประสิทธิ์ความคล้ายแจ็คการ์ดข้อมูลในปริภูมิ 2 มิติ

ระหว่างค่าพารามิเตอร์ a และ b นอกจากนี้เรายังสามารถประยุกต์ใช้สูตรข้างต้นสำหรับหาค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ด ระหว่างข้อมูล 2 ข้อมูล ที่ตัวแปรเชิงตัวเลขทุกตัวแปรมีค่าที่เป็นไปได้มากกว่าหรือเท่ากับ 0 ได้ ภาพ 3.13 แสดงแนวคิดของการวัดความคล้ายด้วยสัมประสิทธิ์ความคล้ายแจ็คการ์ดของข้อมูลในปริภูมิ 2 มิติ

ตัวอย่าง 3.13

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงกลุ่ม 3 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (\text{Male}, \text{Singer}, \text{Married}) \text{ และ } \mathbf{x}_2 = (\text{Male}, \text{Actor}, \text{InRelationship})$$

เมื่อทำการเข้ารหัสข้อมูลทั้งสองด้วยวิธีเข้ารหัสแบบวัน-ฮอท จะได้

$$\mathbf{x}_1 = (1, 0|1, 0, 0|0, 0, 0, 1) \text{ และ } \mathbf{x}_2 = (1, 0|0, 1, 0|0, 0, 1, 0)$$

จะสามารถวัดค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้โดยใช้สมการ 3.22 ดังนี้

$$\begin{aligned} s_J(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\sum_{k=1}^{9} \min(x_{1,k}, x_{2,k})}{\sum_{k=1}^{9} \max(x_{1,k}, x_{2,k})} \\ &= \frac{1+0+0+0+0+0+0+0+0}{1+0+1+1+0+0+0+1+1} \\ &= \frac{1}{5} = 0.2 \end{aligned}$$

ดังนั้น ค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.2

ตัวอย่าง 3.14

พิจารณา ข้อมูล 2 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 4 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (5.1, 3.5, 1.4, 0.2) \text{ และ } \mathbf{x}_2 = (5.0, 3.5, 1.6, 0.6)$$

จะสามารถวัดค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ ได้

ดังนี้

$$
\begin{aligned}
s_J(\mathbf{x}_1, \mathbf{x}_2) &= \frac{\sum_{k=1}^{4} \min(x_{1,k}, x_{2,k})}{\sum_{k=1}^{4} \max(x_{1,k}, x_{2,k})} \\
&= \frac{\min(5.1, 5.0) + \min(3.5, 3.5) + \min(1.4, 1.6) + \min(0.2, 0.6)}{\max(5.1, 5.0) + \max(3.5, 3.5) + \max(1.4, 1.6) + \max(0.2, 0.6)} \\
&= \frac{5.0 + 3.5 + 1.4 + 0.2}{5.1 + 3.5 + 1.6 + 0.6} \\
&= \frac{10.10}{10.80} = 0.94
\end{aligned}
\tag{3.23}
$$

ดังนั้น ค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดระหว่าง $\mathbf{x}_1$ และ $\mathbf{x}_2$ มีค่าเท่ากับ 0.94

เราสามารถแทนระยะห่างระหว่างข้อมูลทั้งหมดในชุดข้อมูลได้ในรูปแบบ เมทริกซ์ระยะทาง (Distance Matrix) หรือ เมทริกซ์ความคล้าย (Similarity Matrix) ซึ่งเป็นเมทริกซ์จตุรัสขนาด $n \times n$ เมื่อ n คือ จำนวนข้อมูลทั้งหมดในชุดข้อมูล และในตำแหน่ง (i, j) ของเมทริกซ์จัดเก็บค่าระยะทาง หรือ ค่าความคล้ายระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ สามารถแสดงเมทริกซ์ระยะทาง หรือ เมทริกซ์ความคล้าย สำหรับชุดข้อมูล $\mathbf{D}$ ได้ดังนี้

$$
\Delta = \begin{bmatrix}
\delta_{x_1,x_1} & \delta_{x_1,x_2} & \cdots & \delta_{x_1,x_n} \\
\delta_{x_2,x_1} & \delta_{x_2,x_2} & \cdots & \delta_{x_2,x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\delta_{x_d,x_1} & \delta_{x_d,x_2} & \cdots & \delta_{x_n,x_n}
\end{bmatrix}
\tag{3.24}
$$

เมื่อ δ_{x_i,x_j} คือ ค่าระยะทางหรือค่าความคล้ายระหว่างข้อมูล $\mathbf{x}_i$ และ $\mathbf{x}_j$ ซึ่งถูกคำนวณด้วยมาตรวัดระยะทางหรือความคล้ายวิธีการใดวิธีการหนึ่ง เราจะเรียกเมทริกซ์ Δ ว่าเมทริกซ์ระยะทาง เมื่อค่าระยะห่างระหว่างข้อมูลถูกคำนวณด้วยมาตรวัดระยะทาง และจะเรียกเมทริกซ์ Δ ว่าเมทริกซ์ความคล้าย เมื่อค่าระยะห่างระหว่างข้อมูลถูกคำนวณด้วยมาตรวัดความคล้าย เมทริกซ์ Δ นี้อาจเป็นเมทริกซ์สมมาตรหรือไม่ก็ได้ ทั้งนี้ขึ้นอยู่กับมาตรวัดระยะห่างระหว่างข้อมูลที่นำมาใช้

การจัดกลุ่มแบบเคมีน

การจัดกลุ่มแบบเคมีน (K-means Clustering) เป็นการแบ่งข้อมูลออกเป็น k กลุ่ม โดยระยะทางระหว่างข้อมูลในกลุ่มเดียวกันมีค่าน้อยกว่าระยะทางไปยังข้อมูลที่อยู่ต่างกลุ่มกัน เราสามารถนิยามปัญหาการจัดกลุ่มข้อมูลดังกล่าวให้อยู่

ในรูปการคำนวณ โดยกำหนดให้ $\mu_j = (\mu_1, \mu_2, ..., \mu_d)$ เป็นจุดศูนย์กลางของกลุ่มข้อมูล C_j บนปริภูมิลักษณะเด่น d มิติ เมื่อ $j = 1, 2, ..., k$ โดย μ_i เป็นค่าเฉลี่ยของตัวแปร X_i ของข้อมูลในกลุ่ม C_j และนิยามฟังก์ชันผลรวมกำลังสองของระยะทางระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มที่ข้อมูลนั้นอยู่ ดังนี้

$$J = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in C_j} d(\mathbf{x}_i, \mu_j)^2 \tag{3.25}$$

เป้าหมายของการจัดกลุ่มแบบเคมีน คือ การแบ่งข้อมูลออกเป็น k กลุ่ม โดยผลรวมกำลังสองของระยะทางระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มของข้อมูลนั้นมีค่าน้อยที่สุด เราสามารถแสดงเป้าหมายดังกล่าวในรูปฟังก์ชันวัตถุประสงค์ (Objective Function) ได้ดังนี้

$$C^* = \arg\min_{C} J \tag{3.26}$$

เมื่อ $C = \{C_1, C_2, ..., C_k\}$ คือ เซตของกลุ่มข้อมูล k กลุ่ม โดย C_j เป็นเซตของข้อมูลที่ถูกจัดอยู่ในกลุ่ม

จากการนิยามปัญหาการจัดกลุ่มแบบเคมีนข้างต้น สามารถแก้ปัญหานี้ได้โดยใช้กระบวนการวนซ้ำ (Iterative Process) มีขั้นตอน ดังนี้

ขั้นตอนที่ 1 สุ่มจุดบนปริภูมิลักษณะเด่นจำนวน k จุด เพื่อแทนจุดศูนย์กลางของข้อมูลแต่ละกลุ่ม จะได้ $\{\mu_1, \mu_2, ..., \mu_k\}$

ขั้นตอนที่ 2 สร้างเซตว่าง C_j สำหรับ $j = 1, 2, ..., k$ เพื่อใช้เก็บข้อมูลที่ถูกจัดอยู่ในกลุ่ม C_j

ขั้นตอนที่ 3 กำหนดตัวนับจำนวนรอบของการทำซ้ำ t ให้มีค่าเริ่มต้นเท่ากับ 0

ขั้นตอนที่ 4 สำหรับข้อมูลแต่ละข้อมูล x_i ในชุดข้อมูล D ทำขั้นตอนดังนี้

4.1 คำนวณค่าระยะทางระหว่าง x_i และ μ_j เมื่อ $j = 1, 2, ..., k$ สำหรับจุดศูนย์กลางของแต่ละกลุ่มข้อมูล

4.2 หากลุ่มข้อมูล C_j ที่ค่าระยะทางระหว่าง x_i และ μ_j มีค่าน้อยที่สุด

4.3 นำข้อมูล x_i เข้าอยู่ในกลุ่ม C_j

ขั้นตอนที่ 5 สำหรับแต่ละกลุ่มข้อมูล C_j เมื่อ $j = 1, 2, ..., k$ ทำปรับค่าจุดศูนย์กลางข้อมูล μ_j โดย $\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathrm{x}_i$

ขั้นตอนที่ 6 เพิ่มค่าตัวนับ t ขึ้น 1

ขั้นตอนที่ 7 กลับไปเริ่มดำเนินการตั้งแต่ขั้นตอนที่ 4 จนกระทั่ง ไม่มีการเปลี่ยนกลุ่มของข้อมูล หรือจำนวนรอบของการทำซ้ำ t มีค่ามากเกินกว่าที่กำหนด

จะเห็นได้ว่าขั้นตอนวิธีการจัดกลุ่มแบบเคมีนนั้น เป็นการดำเนินการวนซ้ำของการจัดข้อมูลแต่ละข้อมูลเข้ากลุ่ม (ในขั้นตอนที่ 4) และการคำนวณหา

จุดศูนย์กลางกลุ่มข้อมูล (ในขั้นตอนที่ 5) กระบวนการวนซ้ำนี้จะถูกดำเนินการต่อเนื่องไปจนกระทั่งไม่มีข้อมูลใดถูกเปลี่ยนกลุ่ม ในบางครั้งอาจต้องใช้จำนวนรอบของการวนซ้ำจำนวนมาก อย่างไรก็ตาม เราสามารถกำหนดจำนวนรอบสูงสุด เพื่อให้การวนซ้ำหยุดการทำงาน แม้ว่ายังมีการเปลี่ยนกลุ่มของข้อมูลอยู่ก็ได้ การใช้งานวิธีการจัดกลุ่มแบบเคมีนเราจะต้องกำหนดค่าจำนวนกลุ่มที่ต้องการ เรียกว่าพารามิเตอร์ และมาตรวัดระยะทางที่เหมาะสม ซึ่งขึ้นอยู่กับสมมติฐานและลักษณะของข้อมูลที่ทำการจัดกลุ่ม

ตัวอย่าง 3.15

จากภาพ 3.14 แสดงการแบ่งข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร คือ X_1 และ X_2 จำนวน 20 ข้อมูล (แสดงข้อมูลในแผนภาพการกระจายซ้ายบนของภาพ 3.14) ออกเป็น 2 กลุ่มด้วยวิธีการจัดกลุ่มแบบเคมีน เมื่อกำหนด $k = 2$ เริ่มต้นด้วยการสุ่มจุดศูนย์กลางกลุ่ม 2 จุด (แทนด้วยสัญลักษณ์ ★ ในแผนภาพการกระจายขวาบน) ได้ค่า ดังนี้

$$\mu_1 = (2.50, 2.05) \text{ และ } \mu_2 = (4.30, 5.07)$$

ในการทำงานรอบที่ 1 ($t = 1$) ทำการจัดข้อมูลแต่ละข้อมูลเข้ากลุ่มที่ระยะห่างจากข้อมูลและจุดศูนย์กลางของกลุ่มมีค่าน้อยที่สุด ได้ผลลัพธ์ดังแผนภาพการกระจายขวาบนของภาพ 3.14 ข้อมูลที่ถูกจัดเข้ากลุ่มที่ 1 แทนด้วยสัญลักษณ์ ◆ และข้อมูลที่ถูกจัดเข้ากลุ่มที่ 2 แทนด้วยสัญลักษณ์ + ต่อมาทำการปรับจุดศูนย์กลางกลุ่ม ได้ค่าดังนี้

$$\mu_1 = (2.24, 1.45) \text{ และ } \mu_2 = (4.02, 4.80)$$

ในรอบการทำงานที่ 2 ($t = 2$) ทำการปรับกลุ่มข้อมูลอีกครั้งโดยนำข้อมูลแต่ละข้อมูลเข้ากลุ่มที่ระยะห่างจากข้อมูลและจุดศูนย์กลางของกลุ่มมีค่าน้อยที่สุด ได้ผลลัพธ์ดังแผนภาพการกระจายซ้ายล่าง แล้วจึงทำการปรับจุดศูนย์กลางกลุ่ม ได้ค่าดังนี้

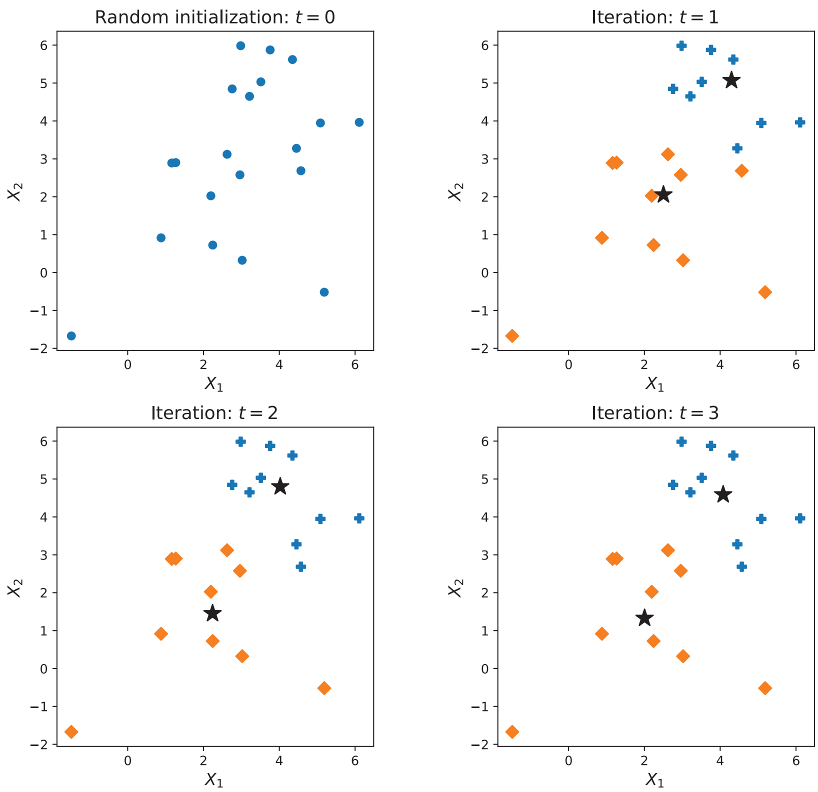
$$\mu_1 = (2.01, 1.33) \text{ และ } \mu_2 = (4.08, 4.59)$$

ในรอบการทำงานที่ 3 ($t = 3$) ทำการจัดกลุ่มข้อมูลอีกครั้ง โดยใช้จุดศูนย์กลางที่ถูกคำนวณค่าใหม่แล้ว ได้ผลลัพธ์ดังแผนภาพการกระจายขวาล่าง แล้วทำการปรับจุดศูนย์กลางกลุ่ม ได้ค่าดังนี้

$$\mu_1 = (2.01, 1.33) \text{ และ } \mu_2 = (4.08, 4.59)$$

จะสังเกตเห็นว่ากลุ่มข้อมูลไม่มีการเปลี่ยนแปลง จึงทำให้ค่าจุดศูนย์กลางกลุ่มที่ถูกคำนวณใหม่ไม่เปลี่ยนแปลงตามไปด้วย ดังนั้น จึงหยุดกระบวนการวนซ้ำเมื่อสิ้นสุดการทำงานในรอบที่ 3 นี้

ผลลัพธ์การจัดกลุ่มข้อมูลแสดงดังแผนภาพการกระจายขวาล่างของภาพ 3.14 โดยข้อมูลถูกจัดเป็น 2 กลุ่ม โดยข้อมูลที่ถูกจัดอยู่ในกลุ่มหนึ่งแทนด้วยสัญลักษณ์ ◆ และข้อมูลที่ถูกจัดอยู่ในอีกกลุ่มหนึ่ง แทนด้วยสัญลักษณ์ +



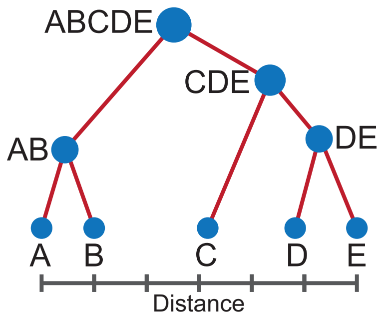
ภาพ 3.14: ตัวอย่างการจัดกลุ่มแบบเคมีนของข้อมูลบนปริภูมิ 2 มิติ โดยแบ่งข้อมูลออกเป็น 2 กลุ่ม แทนจุดข้อมูลในแต่ละกลุ่มด้วยสัญลักษณ์ที่แตกต่างกัน คือ ◆ และ + และจุดศูนย์กลางกลุ่มแทนด้วยสัญลักษณ์ ★

การจัดกลุ่มแบบลำดับชั้น

การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) เป็นการจัดกลุ่มโดยการสร้างลำดับการแบ่งหรือรวมข้อมูลซ้อนๆ กัน ซึ่งสามารถแสดงของลำดับการแบ่งหรือรวมกลุ่มข้อมูลนั้นได้ด้วยแผนภาพเดนโดรแกรม (Dendrogram) ดังแสดงตัวอย่างในภาพ 3.15 โดยโหนดในระดับล่างสุดของต้นไม้หรือโหนดใบ (Leave Node) เป็นกลุ่มข้อมูลขนาดเล็กที่สุดซึ่งบรรจุข้อมูลเพียงข้อมูลเดียว โหนดระหว่างโหนดใบและโหนดราก (Root Node) เป็นกลุ่มข้อมูลที่มีขนาดค่อยๆ เติบโตขึ้น ซึ่งเกิดจากการรวมกลุ่มข้อมูลขนาดเล็กกว่า (ในโหนดระดับล่าง) เข้าด้วยกัน จนกระทั่งที่โหนดรากเป็นกลุ่มข้อมูลที่มีขนาดใหญ่ที่สุดซึ่งรวมทุกๆ ข้อมูลเข้าไปในกลุ่มเดียว หรือในอีกมุมมองหนึ่งคือ กลุ่มข้อมูลในโหนด

ระดับบนค่อยๆ ถูกแบ่งให้มีขนาดเล็กลงเป็นโหนดในระดับล่าง จนกระทั่งกลุ่มข้อมูลเหลือเพียงข้อมูลเดียวเป็นสมาชิกซึ่งไม่สามารถถูกแบ่งได้อีก

การจัดกลุ่มแบบลำดับชั้น สามารถทำได้ 2 แนวทาง คือ 1) การแบ่งข้อมูลจากกลุ่มใหญ่ค่อยๆ แบ่งกลุ่มให้เล็กลงจนได้จำนวนกลุ่มที่ต้องการ เรียกว่า การแบ่งกลุ่มแบบลำดับชั้น (Divisive Hierarchical Clustering) และ 2) การมองข้อมูลแต่ละข้อมูลเป็นกลุ่มข้อมูลขนาดเล็กที่สุด แล้วค่อยๆ รวมกลุ่มขนาดเล็กทีละคู่ให้มีขนาดใหญ่ขึ้น จนกระทั่งได้จำนวนกลุ่มตามที่ต้องการ เรียกว่า การรวมกลุ่มแบบลำดับชั้น (Agglomerative Hierarchical Clustering) ในที่นี้จะกล่าวถึงเฉพาะวิธีการรวมกลุ่มแบบลำดับชั้นเท่านั้น



ภาพ 3.15: ตัวอย่างแผนภาพเดนโดรแกรมแสดงการแบ่งกลุ่มแบบลำดับชั้นของข้อมูล A B C D และ E

ตัวอย่าง 3.16

พิจารณาภาพ 3.15 ซึ่งแสดงตัวอย่างการจัดกลุ่มแบบลำดับชั้นของข้อมูล $\{A, B, C, D, E\}$ แผนภาพเดนโดรแกรมดังกล่าวแสดงลำดับการจัดกลุ่มด้วยวิธีการรวมกลุ่มแบบลำดับชั้น ดังนี้

ลำดับ	กลุ่มข้อมูล
1	$\{A\}, \{B\}, \{C\}, \{D\}, \{E\}$
2	$\{A, B\}, \{C\}, \{D\}, \{E\}$
3	$\{A, B\}, \{C\}, \{D, E\}$
4	$\{A, B\}, \{C, D, E\}$
5	$\{A, B, C, D, E\}$

หากต้องการแบ่งข้อมูลออกเป็น 2 กลุ่ม กระบวนการทำงานจะหยุดเมื่อเสร็จสิ้นการทำงานในลำดับที่ 4 และจะได้กลุ่มข้อมูลผลลัพธ์ คือ $\{A, B\}$ และ $\{C, D, E\}$

เมื่อกำหนดจำนวนกลุ่มข้อมูลที่ต้องการ k วิธีการรวมกลุ่มแบบลำดับชั้น บนชุดข้อมูล D มีขั้นตอนดังนี้

ขั้นตอนที่ 1 สร้างเซตของกลุ่มข้อมูล C สำหรับแต่ละข้อมูล x_i ในชุดข้อมูล D จะได้ $C = \{C_1, C_2, ..., C_n\}$ เมื่อ $C_i = \{\mathrm{x}_i\}$ สำหรับ $i = 1, 2, ..., n$ และ n คือจำนวนข้อมูลในชุดข้อมูล D

ขั้นตอนที่ 2 คำนวณเมทริกซ์ระยะทาง Δ สำหรับชุดข้อมูล D โดยใช้สมการ (3.24)

ขั้นตอนที่ 3 คำนวณหาค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j แทนด้วยสัญลักษณ์ δ_{C_i, C_j} สำหรับทุกๆ คู่ของกลุ่มข้อมูลในเซต C โดยอาศัยเมทริกซ์ระยะทาง Δ

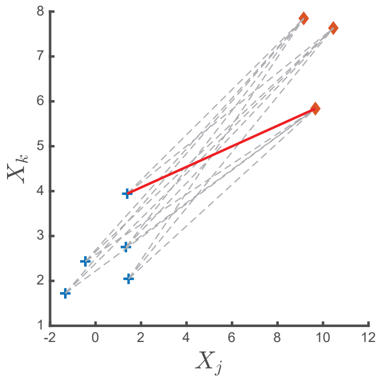
ขั้นตอนที่ 4 หาคู่ของกลุ่มข้อมูล C_i และ C_j ที่มีค่าระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด

ขั้นตอนที่ 5 สร้างกลุ่มข้อมูลใหม่ C_{ij} โดยการรวมกลุ่มข้อมูล C_i และ C_j เข้าด้วยกัน

ขั้นตอนที่ 6 ทำการลบกลุ่มข้อมูล C_i และ C_j ออกจากเซต C และเพิ่มข้อมูลใหม่ C_{ij} เข้าไปยังเซต C

ขั้นตอนที่ 7 กลับไปเริ่มการดำเนินการตั้งแต่ขั้นตอนที่ 3 จนกระทั่งจำนวนสมาชิกในเซต C มีค่าเท่ากับ k

จากขั้นตอนวิธีการแบ่งกลุ่มแบบลำดับชั้นในขั้นตอนที่ 3 เราจำเป็นต้องคำนวณค่าระยะทางระหว่างกลุ่มข้อมูล 2 กลุ่มข้อมูลใดๆ กำหนดให้ C_i และ C_j เป็นกลุ่มข้อมูล 2 กลุ่ม ระยะทางระหว่างกลุ่มข้อมูลทั้งสอง แทนด้วยสัญลักษณ์ δ_{C_i,C_j} สามารถคำนวณได้ด้วยวิธีการ ต่อไปนี้

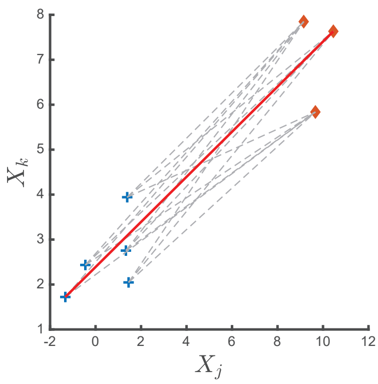


ภาพ 3.16: ตัวอย่างการเชื่อมต่อเดี่ยวระหว่างกลุ่มข้อมูล 2 กลุ่มบนปริภูมิ 2 มิติ โดยจุดข้อมูลในแต่ละกลุ่มแทนด้วยสัญลักษณ์ที่แตกต่างกัน คือ ◆ และ + เส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลจากต่างกลุ่มข้อมูล เป็นระยะทางที่น้อยที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง ซึ่งเชื่อมต่อกลุ่มข้อมูลเข้าด้วยกัน

การเชื่อมต่อเดี่ยว (Single-Linkage) เป็นการใช้ค่าระยะทางที่น้อยที่สุดระหว่างข้อมูล 2 ข้อมูล โดย ข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_i และอีกข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_j เป็นค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j สามารถเขียนในรูปสมการคณิตศาสตร์ได้ดังนี้

$$\delta_{C_i,C_j} = \min\{\delta_{\mathrm{x}_p,\mathrm{x}_q} | \mathrm{x}_p \in C_i, \mathrm{x}_q \in C_j\} \tag{3.27}$$

จะสังเกตได้ว่ากลุ่มข้อมูลทั้งสอง เชื่อมต่อกันด้วยข้อมูล 2 ข้อมูลจากคนละกลุ่มที่อยู่ใกล้กันที่สุดเพียง 1 คู่ (1 การเชื่อมต่อ) เท่านั้น แสดงตัวอย่างการเชื่อมต่อเดี่ยวระหว่างกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ ดังภาพ 3.16 โดยเส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลจากกลุ่มข้อมูลต่างกัน เป็นระยะทางที่น้อยที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง



ภาพ 3.17: ตัวอย่างการเชื่อมต่อสมบูรณ์ระหว่างกลุ่มข้อมูล 2 กลุ่ม โดยจุดข้อมูลในแต่ละกลุ่มแทนด้วยสัญลักษณ์ที่แตกต่างกัน คือ ◆ และ + เส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลจากต่างกลุ่มข้อมูลกัน เป็นระยะทางที่มากที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง ซึ่งทำให้เกิดการเชื่อมต่อกันระหว่างกลุ่มข้อมูล 2 กลุ่มอย่างสมบูรณ์

การเชื่อมต่อสมบูรณ์ (Complete-Linkage) เป็นการใช้ค่าระยะทางที่มากที่สุดระหว่างข้อมูล 2 ข้อมูล โดย ข้อมูลหนึ่งมาจากกลุ่มข้อมูล C_i และอีกข้อมูลหนึ่งมาจากกลุ่มข้อมูล C_j เป็นค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

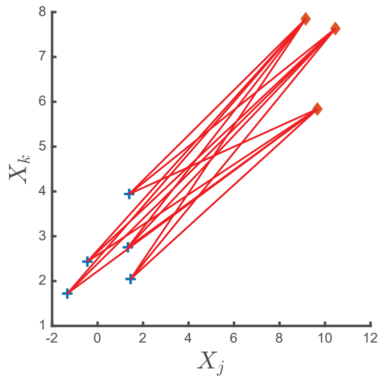
$$\delta_{C_i,C_j} = \max\{\delta_{\mathrm{x}_p,\mathrm{x}_q} | \mathrm{x}_p \in C_i, \mathrm{x}_q \in C_j\} \tag{3.28}$$

แม้ว่าการวัดระยะทางระหว่างกลุ่มข้อมูลวิธีนี้จะใช้ค่าระยะทางที่มากที่สุดระหว่างข้อมูล 2 ข้อมูลที่มาจากคนละกลุ่มข้อมูลเพียงค่าเดียว เมื่อคู่ของข้อมูลจากทั้ง 2 กลุ่มข้อมูลที่อยู่ห่างกันมากที่สุดสร้างการเชื่อมต่อกลุ่มข้อมูลทั้ง 2 กลุ่ม จึงทำให้คู่ข้อมูลอื่นๆ ที่เป็นไปได้ทั้งหมด ระหว่างกลุ่มข้อมูลทั้ง 2 กลุ่ม อยู่ภาย

ใต้ระยะทางนี้ด้วย และทำให้กลุ่มข้อมูลทั้งสองเชื่อมต่อกันอย่างสมบูรณ์ ภาพ 3.17 แสดงตัวอย่างการเชื่อมต่อสมบูรณ์ระหว่างกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ เส้นทึบในภาพแสดงระยะทางที่มากที่สุดระหว่างข้อมูลจากกลุ่มข้อมูลทั้งสอง

การเชื่อมต่อเฉลี่ย (Average-Linkage) ค่าระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j เป็นค่าเฉลี่ยของระยะทางระหว่างข้อมูลทุกคู่ที่เป็นไปได้ โดยข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_i และอีกข้อมูลหนึ่งเป็นสมาชิกของกลุ่มข้อมูล C_j สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$\delta_{C_i,C_j} = \frac{\sum_{x_p \in C_i} \sum_{x_q \in C_j} \delta_{x_p,x_q}}{|C_i| \cdot |C_j|} \tag{3.29}$$

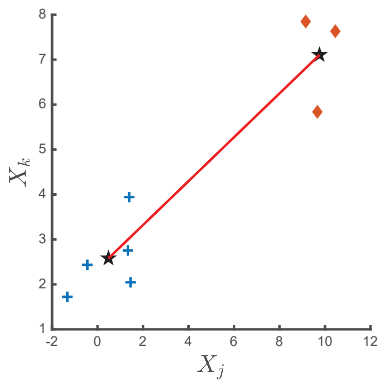


ภาพ 3.18: ตัวอย่างการเชื่อมต่อเฉลี่ยระหว่างกลุ่มข้อมูล 2 กลุ่ม โดยจุดข้อมูลในแต่ละกลุ่มแทนด้วยสัญลักษณ์ที่แตกต่างกัน คือ ◆ และ + เส้นทึบที่เชื่อมระหว่างข้อมูล 2 ข้อมูลทุกๆ คู่ที่เป็นไปได้ โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ถูกใช้ในการคำนวณระยะทางของการเชื่อมต่อเฉลี่ย

ภาพ 3.18 แสดงการเชื่อมโยงข้อมูลทุกๆ คู่ที่เป็นไปได้โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ระยะทางของคู่ข้อมูลแต่ละคู่นี้ถูกนำมาเฉลี่ยเพื่อใช้แทนค่าระยะทางระหว่างกลุ่มข้อมูลทั้ง 2 กลุ่ม

ระยะทางระหว่างศูนย์กลางกลุ่ม (Centroid Distance) เป็นการวัดระยะทางระหว่างกลุ่มข้อมูลโดยใช้จุดศูนย์กลางของกลุ่มข้อมูลเป็นตัวแทนของข้อมูลในกลุ่ม และวัดระยะทางระหว่างจุดศูนย์กลางของกลุ่มทั้งสอง กำหนดให้ μ_i และ μ_j เป็นจุดศูนย์กลางของกลุ่ม C_i และ C_j ตามลำดับ ระยะทางระหว่างกลุ่มข้อมูล C_i และ C_j คือ

$$\delta_{C_i,C_j} = \delta_{\mu_i,\mu_j} \tag{3.30}$$



ภาพ 3.19: ตัวอย่าง ระยะ ทาง ระหว่างศูนย์กลาง กลุ่ม โดย จุด ข้อมูล ใน แต่ละ กลุ่มแทนด้วยสัญลักษณ์ที่แตกต่างกัน คือ ◆ และ + และจุดศูนย์กลางกลุ่มแทนด้วยสัญลักษณ์ ★ เส้นทึบเชื่อมระหว่างจุดศูนย์กลางกลุ่มทั้งสองซึ่งใช้แทนระยะทางระหว่างกลุ่มข้อมูล

ภาพ 3.19 แสดงตัวอย่างระยะทางระหว่างศูนย์กลางกลุ่ม ของกลุ่มข้อมูล 2 กลุ่ม บนปริภูมิ 2 มิติ โดยจุดศูนย์กลางกลุ่มข้อมูลแต่ละกลุ่ม แทนด้วยสัญลักษณ์ ★

ตัวอย่าง 3.17

พิจารณาเมทริกซ์ระยะทางของข้อมูล 5 ข้อมูล ต่อไปนี้

$$\Delta = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0 & 2.40 & 3.27 & 10.45 & 11.68 \\ 2.40 & 0 & 0.93 & 8.21 & 9.33 \\ 3.27 & 0.93 & 0 & 7.27 & 8.42 \\ 10.45 & 8.21 & 7.27 & 0 & 1.86 \\ 11.68 & 9.33 & 8.42 & 1.86 & 0 \end{bmatrix} \end{array}$$

กำหนดกลุ่มข้อมูล C_1 และ C_2 โดย $C_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ และ $C_2 = \{\mathbf{x}_4, \mathbf{x}_5\}$ จะสามารถหาระยะทางระหว่างกลุ่มข้อมูลทั้ง 2 ด้วยวิธีต่างๆ ได้ดังนี้

วิธีการเชื่อมต่อเดี่ยว พิจารณาหาค่าระยะทางที่น้อยที่สุดจากค่าระยะทางของคู่ข้อมูลที่เป็นไปได้ทั้งหมด โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ดังนี้

$$\begin{aligned}\delta_{C_i,C_j} &= \min\{\delta_{\mathrm{x}_1,\mathrm{x}_4}, \delta_{\mathrm{x}_1,\mathrm{x}_5}, \delta_{\mathrm{x}_2,\mathrm{x}_4}, \delta_{\mathrm{x}_2,\mathrm{x}_5}, \delta_{\mathrm{x}_3,\mathrm{x}_4}, \delta_{\mathrm{x}_3,\mathrm{x}_5}\}\\ &= \min\{10.45, 11.68, 8.21, 9.33, 7.27, 8.42\}\\ &= 7.27\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อวัดด้วยวิธีการเชื่อมต่อเดี่ยว มีค่าเท่ากับ 7.27

วิธีการเชื่อมต่อสมบูรณ์ พิจารณาหาค่าระยะทางที่มากที่สุดจากค่าระยะทางของคู่ข้อมูลที่เป็นไปได้ทั้งหมด โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ดังนี้

$$\begin{aligned}\delta_{C_i,C_j} &= \max\{\delta_{\mathrm{x}_1,\mathrm{x}_4}, \delta_{\mathrm{x}_1,\mathrm{x}_5}, \delta_{\mathrm{x}_2,\mathrm{x}_4}, \delta_{\mathrm{x}_2,\mathrm{x}_5}, \delta_{\mathrm{x}_3,\mathrm{x}_4}, \delta_{\mathrm{x}_3,\mathrm{x}_5}\}\\ &= \max\{10.45, 11.68, 8.21, 9.33, 7.27, 8.42\}\\ &= 11.68\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อวัดด้วยวิธีการเชื่อมต่อสมบูรณ์ มีค่าเท่ากับ 11.68

วิธีการเชื่อมต่อเฉลี่ย หาค่าเฉลี่ยของค่าระยะทางของคู่ข้อมูลที่เป็นไปได้ทั้งหมด โดยข้อมูลแต่ละคู่มาจากต่างกลุ่มข้อมูลกัน ดังนี้

$$\begin{aligned}\delta_{C_i,C_j} &= \frac{\delta_{\mathrm{x}_1+\mathrm{x}_4} + \delta_{\mathrm{x}_1,\mathrm{x}_5} + \delta_{\mathrm{x}_2,\mathrm{x}_4} + \delta_{\mathrm{x}_2,\mathrm{x}_5} + \delta_{\mathrm{x}_3,\mathrm{x}_4} + \delta_{\mathrm{x}_3,\mathrm{x}_5}}{6}\\ &= \frac{10.45 + 11.68 + 8.21 + 9.33 + 7.27 + 8.42}{6}\\ &= \frac{55.36}{6} = 9.26\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อวัดด้วยวิธีการเชื่อมต่อเฉลี่ย มีค่าเท่ากับ 9.26

ระยะทางระหว่างศูนย์กลางกลุ่ม เมื่อคำนวณค่าศูนย์กลางกลุ่มของกลุ่มข้อมูล C_1 และ C_2 ได้ ดังนี้

$$\mu_1 = (1.38, 3.83) \text{ และ } \mu_2 = (10.00, 7.05)$$

จะสามารถคำนวณระยะทางระหว่างศูนย์กลางกลุ่มทั้งสองด้วยระยะทางแบบยุคลิด ได้ดังนี้

$$\begin{aligned}\delta_{C_i,C_j} &= \sqrt{(1.38 - 10.00)^2 + (3.83 - 7.05)^2} \\ &= \sqrt{(-8.62)^2 + (-3.22)^2} \\ &= \sqrt{74.30 + 10.37} \\ &= \sqrt{84.67} = 9.20\end{aligned}$$

ดังนั้น ระยะทางระหว่างกลุ่มข้อมูล C_1 และ C_2 เมื่อใช้ระยะทางระหว่างศูนย์กลางกลุ่ม มีค่าเท่ากับ 9.20

ตัวอย่าง 3.18

จากข้อมูล 5 ข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร ต่อไปนี้

$$\mathbf{x}_1 = (0.80, 4.12) \quad \mathbf{x}_2 = (0.88, 4.17) \quad \mathbf{x}_3 = (2.49, 4.33)$$
$$\mathbf{x}_4 = (8.79, 8.36) \text{ และ } \mathbf{x}_5 = (10.72, 7.62)$$

สามารถจัดกลุ่มข้อมูลทั้ง 5 ข้อมูลออกเป็น 2 กลุ่ม ด้วยวิธีการรวมกลุ่มแบบลำดับชั้น ได้ดังนี้

เริ่มต้นด้วยการสร้างเซตของกลุ่มข้อมูล $C = \{C_1, C_2, C_3, C_4, C_5\}$ สำหรับแต่ละข้อมูล ต่อมาทำการคำนวณเมทริกซ์ระยะทางระหว่างข้อมูล Δ ได้ดังนี้

$$\Delta = \begin{array}{c} \\ \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \mathbf{x}_4 \\ \mathbf{x}_5 \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4 & \mathbf{x}_5 \end{array} \\ \begin{bmatrix} 0 & 0.09 & 1.71 & 9.05 & 10.52 \\ 0.09 & 0 & 1.62 & 8.96 & 10.43 \\ 1.71 & 1.62 & 0 & 7.48 & 8.86 \\ 9.05 & 8.96 & 7.48 & 0 & 2.06 \\ 10.52 & 10.43 & 8.86 & 2.06 & 0 \end{bmatrix} \end{array}$$

ตัวอย่างนี้ใช้วิธีการเชื่อมต่อสมบูรณ์ในการหาค่าระยะทางระหว่างกลุ่มข้อมูล ทำการหาคู่ของกลุ่มข้อมูลที่มีระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด ได้คู่ของกลุ่มข้อมูล C_1 และ C_2 ซึ่งมีระยะทางระหว่างกลุ่มข้อมูลเท่ากับ 0.09 ทำการรวมกลุ่มข้อมูล C_1 และ C_2 เข้าด้วยกัน ได้กลุ่มข้อมูลใหม่ $C_{12} = \{\mathbf{x}_1, \mathbf{x}_2\}$ และปรับเซตของกลุ่มข้อมูล C ใหม่ ได้ดังนี้

$$C = \{C_{12}, C_3, C_4, C_5\}$$

ต่อมาทำการหาคู่ของกลุ่มข้อมูลที่มีระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด ได้คู่ของกลุ่มข้อมูล C_{12} และ C_3 ซึ่งมีระยะทางระหว่างกลุ่มข้อมูลเท่ากับ 1.71 จึงทำการรวมกลุ่มข้อมูล C_{12} และ C_3 เข้าด้วยกัน ได้กลุ่มข้อมูลใหม่ $C_{123} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ และปรับเซตของกลุ่มข้อมูล C ใหม่ ได้ดังนี้

$$C = \{C_{123}, C_4, C_5\}$$

ต่อมาทำการหาคู่ของกลุ่มข้อมูลที่มีระยะทางระหว่างกลุ่มข้อมูลน้อยที่สุด ได้คู่ของกลุ่มข้อมูล C_4 และ C_5 ซึ่งมีระยะทางระหว่างกลุ่มข้อมูลเท่ากับ 2.06 จึงทำการรวมกลุ่มข้อมูล C_4 และ C_5 เข้าด้วยกัน ได้กลุ่มข้อมูลใหม่ $C_{45} = \{\mathbf{x}_4, \mathbf{x}_5\}$ และปรับเซตของกลุ่มข้อมูล C ใหม่ ได้ดังนี้

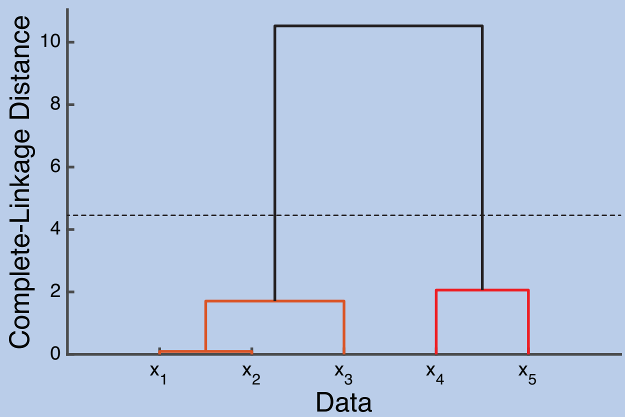
$$C = \{C_{123}, C_{45}\}$$

จากผลลัพธ์พบว่าจำนวนกลุ่มข้อมูลที่ได้มีค่าเท่ากับจำนวนกลุ่มข้อมูลที่ต้องการ จึงหยุดกระบวนการรวมกลุ่มข้อมูล

ดังนั้น การจัดกลุ่มข้อมูล 5 ข้อมูลข้างต้น ออกเป็น 2 กลุ่ม ด้วยวิธีการรวมกลุ่มแบบลำดับชั้น ได้ผลลัพธ์ คือ

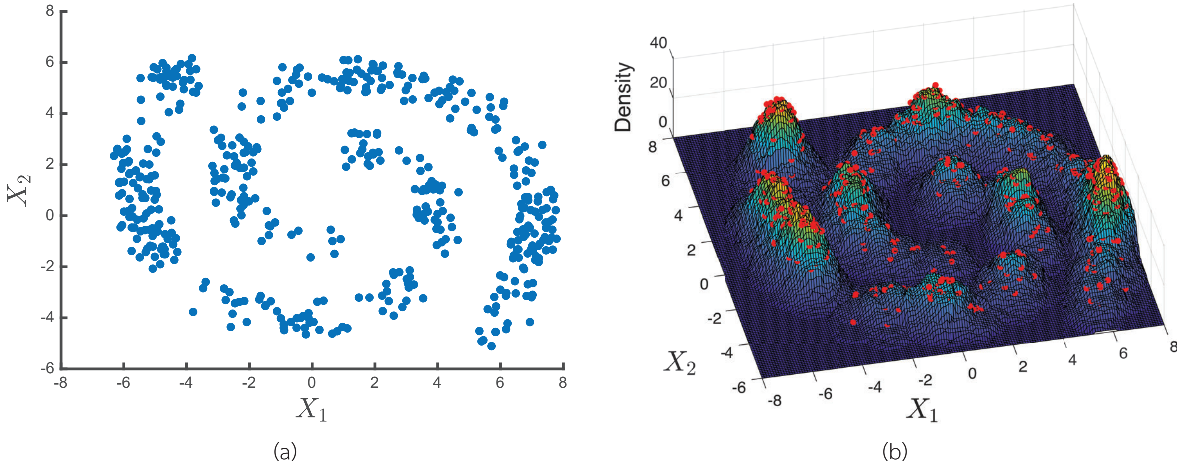
$$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\} \text{ และ } \{\mathbf{x}_4, \mathbf{x}_5\}$$

และสามารถแสดงลำดับการแบ่งกลุ่มข้อมูลแบบลำดับชั้นด้วยแผนภาพเดนโดรแกรม ได้ดังนี้



การจัดกลุ่มบนฐานความหนาแน่น

การจัดกลุ่มบนฐานความหนาแน่น (Density-Based Clustering) เป็นการใช้ความหนาแน่นของข้อมูลในการพิจารณาจัดกลุ่มข้อมูล วิธีการจัดกลุ่มบน

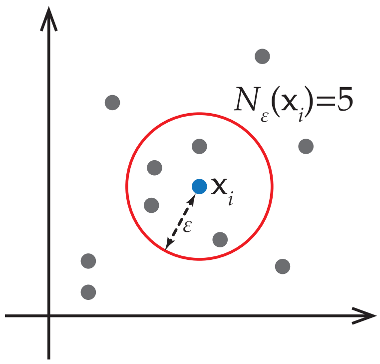


ภาพ 3.21: มุมมองการจัดกลุ่มด้วยวิธีดีบีสแกน (a) แผนภาพการกระจายของข้อมูลบนปริภูมิ 2 มิติ (b) แผนภาพพื้นผิวแสดงความหนาแน่นของข้อมูล

ฐานความหนาแน่นโดยส่วนใหญ่ทำให้ได้กลุ่มข้อมูลผลลัพธ์ที่ไม่ถูกกำจัดด้วยสมมติฐานของรูปร่างของกลุ่มข้อมูล อีกทั้งยังสามารถจัดการกับข้อมูลผิดปรกติ (Outlier) ได้อย่างมีประสิทธิภาพ วิธีการหนึ่งในการจัดกลุ่มข้อมูลบนฐานความหนาแน่นที่เป็นที่รู้จัก คือ วิธีดีบีสแกน (Density-Based Spatial Clustering of Applications with Noise: DBSCAN) โดยวิธีการนี้กำหนดนิยามความหนาแน่นของข้อมูล ดังนี้ กำหนดให้ $V_\varepsilon(\mathrm{x}_i)$ เป็นเซตของจุดข้อมูลในชุดข้อมูล D ที่อยู่ภายในทรงกลม ขนาดรัศมี ε ซึ่งเป็นพารามิเตอร์ที่ต้องถูกกำหนดค่า โดยมีจุดข้อมูล x_i เป็นจุดศูนย์กลาง สามารถเขียนในรูปสัญลักษณ์คณิตศาสตร์ ได้ดังนี้

$$V_\varepsilon(\mathrm{x}_i) = \{\mathrm{p} \in \mathrm{D} | \delta_{\mathrm{x}_i,\mathrm{p}} \le \varepsilon\} \tag{3.31}$$

ค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i แทนด้วยสัญลักษณ์ $N_\varepsilon(\mathrm{x}_i)$ คือ จำนวนข้อมูลที่อยู่ห่างจากข้อมูล x_i ไม่เกินค่ารัศมี ε ทั้งนี้นับรวมข้อมูล x_i ด้วย นั่นคือ $N_\varepsilon(\mathrm{x}_i) = |V_\varepsilon(\mathrm{x}_i)|$ แสดงตัวอย่างการคำนวณค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i ดังภาพ 3.20 ซึ่งภายในวงกลมขนาดรัศมี ε ที่มีจุด x_i เป็นจุดศูนย์กลาง มีจำนวนจุดข้อมูลทั้งหมด 5 จุด ดังนั้น ค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i จึงมีค่าเท่ากับ 5

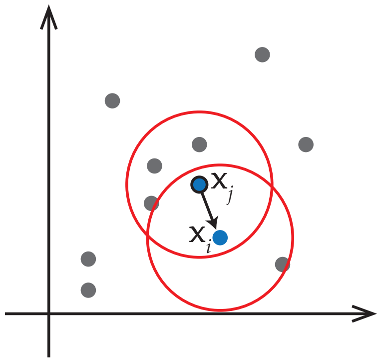


ภาพ 3.20: ตัวอย่างการหาค่าความหนาแน่นของข้อมูล ณ ตำแหน่งจุดข้อมูล x_i เมื่อกำหนดค่ารัศมีวงกลมขนาด ε ทำให้ความหนาแน่นของข้อมูล ณ ตำแหน่งจุดข้อมูล x_i มีค่าเท่ากับ 5

เพื่อให้เข้าใจหลักการของวิธีดีบีสแกนในการจัดกลุ่มข้อมูล ลองจินตนาการว่าจุดข้อมูลถูกวางบนระนาบ 2 มิติที่สอดคล้องกับปริภูมิลักษณะเด่นของข้อมูล ต่อมาทำการยกข้อมูลขึ้นจากระนาบ โดยใช้ค่าความหนาแน่นของข้อมูล ณ ตำแหน่งของแต่ละข้อมูลแทนความสูง เราจะได้ภาพจากจินตนาการ ดังแสดงตัวอย่างในภาพ 3.21(b) ซึ่งมีลักษณะคล้ายกับแผนที่ภูมิประเภทของเทือกเขา วิธีดีบีสแกนทำการจัดกลุ่มข้อมูลที่อยู่บนเทือกเขาเดียวกันเข้าไว้ด้วยกัน ส่วนข้อมูลที่กระจัดกระจายบนพื้นราบ (มีความหนาแน่นต่ำ) นั่น จะถูกจัดเป็นข้อมูลผิดปรกติ

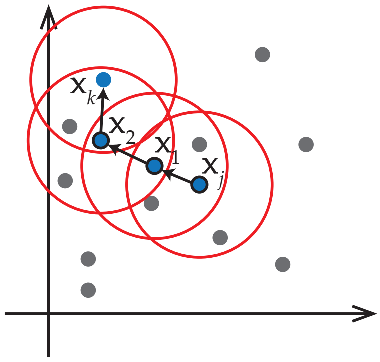
ในการจัดกลุ่มด้วยวิธีดีบีสแกน จุดข้อมูลแต่ละจุดจะถูกจำแนกออกเป็น 3 ประเภท คือ

- **จุดแกน (Core Point)** คือ ข้อมูลที่ความหนาแน่น ณ จุดข้อมูลนั้นมีค่ามากกว่าหรือเท่ากับ ค่า $minPts$ นั่นคือ จุดข้อมูล x_i จะเป็นจุดแกน ก็ต่อเมื่อ $N_\varepsilon(x_i) \geq minPts$ เมื่อ $minPts$ เป็นค่าขีดแบ่งความหนาแน่นเฉพาะถิ่นที่จะต้องถูกกำหนดขึ้น ซึ่งถือว่าเป็นค่าพารามิเตอร์อีกค่าหนึ่งของวิธีดีบีสแกน
- **จุดขอบ (Border Point)** คือ ข้อมูลที่ความหนาแน่น ณ จุดข้อมูลนั้นมีค่าน้อยกว่า ค่า $minPts$ และ เป็นจุดข้อมูลที่อยู่ภายในรัศมี ε ของข้อมูลที่เป็นจุดแกน ข้อมูล x_i จะถูกจัดเป็นจุดขอบ ก็ต่อเมื่อ $N_\varepsilon(x_i) < minPts$ และ $x_i \in V_\varepsilon(x_j)$ เมื่อ x_j เป็นจุดแกนจุดหนึ่ง
- **จุดข้อมูลรบกวน (Noise Point) หรือ ข้อมูลผิดปรกติ** คือ ข้อมูลที่ความหนาแน่น ณ จุดข้อมูลนั้นเป็นจุดศูนย์กลางมีค่าน้อยกว่า ค่า $minPts$ และ ไม่เป็นจุดข้อมูลที่อยู่ภายในรัศมี ε ของข้อมูลที่เป็นจุดแกน นั่นคือ ข้อมูล x_i จะถูกจัดเป็นจุดข้อมูลรบกวน ก็ต่อเมื่อ x_i ไม่เป็นจุดแกนและจุดขอบ



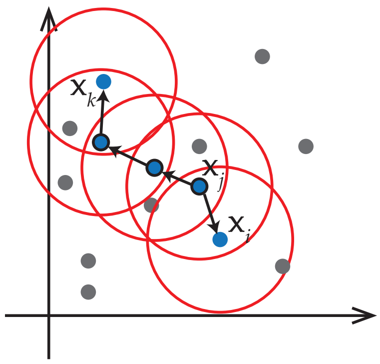
ภาพ 3.22: จุดข้อมูล x_i สามารถไปถึงได้ด้วยความหนาแน่นโดยตรง (Directly Density Reachable) จากจุดข้อมูล x_j เมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$

เราจะกล่าวว่า จุดข้อมูล x_i สามารถไปถึงได้ด้วยความหนาแน่นโดยตรง (Directly Density Reachable) จากจุดข้อมูล x_j ถ้าจุดข้อมูล x_i อยู่ในวงรัศมีขนาด ε ของจุดข้อมูล x_j และ จุดข้อมูล x_j เป็นจุดแกน (แสดงภาพประกอบคำอธิบายดังภาพ 3.22) เสมือนว่าจากจุด x_j เราสามารถลื่นไถลไปยังจุดข้อมูล x_i โดยตรงจากจุดข้อมูล x_j โดยอาศัยความลาดชัดของความหนาแน่นได้ และจุดข้อมูล x_i เป็นจุดข้อมูลในแนวกลุ่มข้อมูล (เทือกเขา) เดียวกับจุดข้อมูล x_j โดยอยู่ในระดับความหนาแน่น (ความสูงจากพื้นราบ) ที่ต่ำกว่า x_j นั่นเอง นอกจากนี้ เราจะกล่าวว่า จุดข้อมูล x_k สามารถไปถึงได้ด้วยความหนาแน่น (Density Reachable) จากจุดข้อมูล x_j ถ้ามีลูกโซ่ของจุดข้อมูล $x_0, x_1, ..., x_l$ ที่ $x_0 = x_k$ และ $x_l = x_j$ และ x_p สามารถไปถึงได้ด้วยความหนาแน่นโดยตรงจาก จุดข้อมูล x_{p-1} สำหรับ $p = 1, 2, ..., l$ (แสดงภาพประกอบคำอธิบายดังภาพ 3.23) กล่าวคือ เราสามารถลื่นไถลจุดแกน x_j ไปยัง x_k โดยผ่านจุดแกน $x_1, x_2, ..., x_{l-1}$ ตามลำดับ

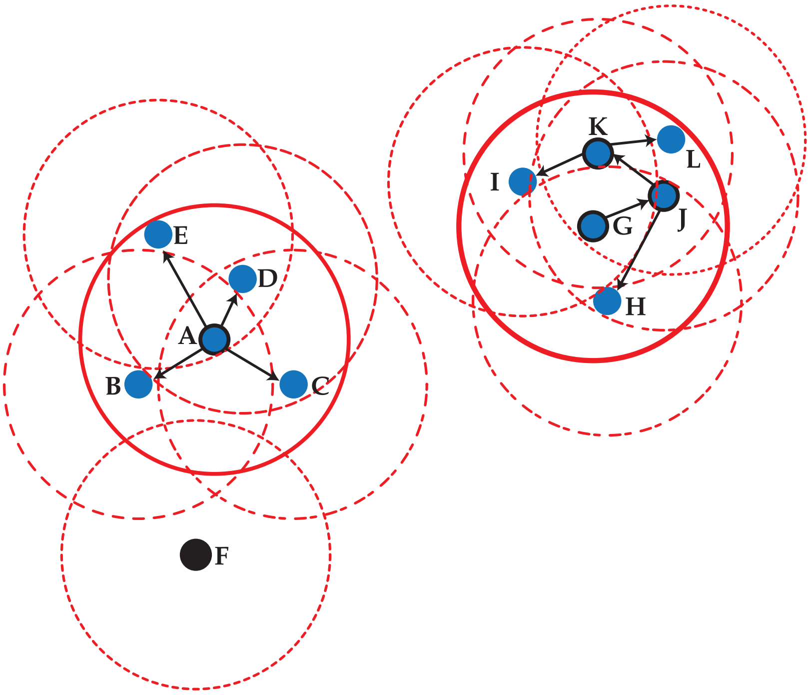


ภาพ 3.23: จุดข้อมูล x_k สามารถไปถึงได้ด้วยความหนาแน่น (Density Reachable) จากจุดข้อมูล x_j ผ่านจุดข้อมูล x_1 และ x_2 ตามลำดับเมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$

กำหนดให้ จุดข้อมูล x_i และ x_k ถูกเชื่อมโยงกันด้วยความหนาแน่น (Density Connected) ถ้ามีจุดแกน x_j ที่ทั้ง x_i และ x_k สามารถไปถึงได้ด้วยความหนาแน่นจากจุดข้อมูล x_j (ดูภาพ 3.24 ประกอบคำอธิบาย) กลุ่มข้อมูลบนฐานความหนาแน่น คือ เซตขนาดใหญ่ที่สุดของจุดข้อมูลที่แต่ละคู่ของจุดข้อมูลที่เป็นไปได้ทั้งหมดในเซตถูกเชื่อมโยงกันด้วยความหนาแน่น วิธีดีบีสแกนมีขั้นตอนดำเนินการหากลุ่มข้อมูลบนฐานความหนาแน่นในชุดข้อมูล D เมื่อกำหนดค่าพารามิเตอร์ ε และ $minPts$ ดังนี้



ภาพ 3.24: จุดข้อมูล x_i และ x_k ถูกเชื่อมโยงกันด้วยความหนาแน่น (Density Connected) โดยมีจุดแกน x_j เป็นจุดเริ่มร่วม เมื่อกำหนดค่าพารามิเตอร์ $minPts = 5$



ภาพ 3.25: ตัวอย่างการจัดกลุ่มด้วยวิธีดีบีสแกนสำหรับข้อมูล 12 ข้อมูล บนปริภูมิ 2 มิติ จุดกลมขอบแข็ง แทน ข้อมูลที่เป็นจุดแกน (Core Point) จุดกลมไม่มีขอบ แทน ข้อมูลที่เป็นจุดขอบ (Border Point) และจุดกลมทึบ แทน ข้อมูลผิดปรกติ (Outlier) เมื่อกำหนดค่าพารามิเตอร์ $minPts$ = 5 จุดข้อมูลแต่ละจุดถูกล้อมด้วยวงกลมรัศมี ε

ขั้นตอนที่ 1 สำหรับทุกๆ ข้อมูล x_i ในข้อมูล D ทำการหาเซต $V_\varepsilon(x_i)$ ซึ่งเป็นเซตของจุดข้อมูลที่อยู่ภายในทรงกลมรัศมี ε ที่มีจุด x_i เป็นจุดศูนย์กลาง

ขั้นตอนที่ 2 คำนวณหาค่าความหนาแน่น ณ ตำแหน่งของจุดข้อมูล x_i ทุกๆ ข้อมูล โดยค่าความหนาแน่น $N_\varepsilon(x_i) = |V_\varepsilon(x_i)|$

ขั้นตอนที่ 3 หาเซตของจุดแกน $Core$ โดย $Core = \{x_i | N_\varepsilon(x_i) \geq minPts\}$

ขั้นตอนที่ 4 กำหนดให้ k มีค่าเท่ากับ 0 สำหรับใช้นับจำนวนกลุ่มข้อมูล

ขั้นตอนที่ 5 กำหนดให้ C เป็นเซตว่าง สำหรับบรรจุกลุ่มข้อมูล

ขั้นตอนที่ 6 กำหนดให้ L เป็นเซตของข้อมูลทั้งหมดในชุดข้อมูล D

ขั้นตอนที่ 7 วนทำซ้ำขั้นตอนต่อไปนี้ จนกระทั่ง เซต $Core$ เป็นเซตว่าง

7.1 นำสมาชิก x_i จากเซต $Core$ มาพิจารณา

7.2 เพิ่มค่า k ขึ้น 1

7.3 สร้างเซต C_k โดย $C_k = \{x_i\}$

7.4 หาจุดข้อมูลทุกจุดข้อมูลในเซต L ที่สามารถไปถึงได้ด้วยความหนาแน่นจากจุดข้อมูล x_i แล้วนำไปเพิ่มเป็นสมาชิกในเซต C_k

7.5 ลบข้อมูลที่เป็นสมาชิกในเซต C_k ออกจากเซต L และเซต $Core$

7.6 นำเซต C_k เพิ่มเป็นสมาชิกของเซต C

ขั้นตอนที่ 8 สร้างเซต $C_{noise} = L$

ขั้นตอนที่ 9 นำเซต C_{noise} เพิ่มเป็นสมาชิกของเซต C

ขั้นตอนที่ 10 ผลลัพธ์ที่ได้ คือ เซต $C = \{C_1, C_2, ...C_k, C_{noise}\}$ ซึ่งเป็นเซตของกลุ่มข้อมูล k กลุ่มและกลุ่มข้อมูลที่ถูกจัดให้เป็นข้อมูลผิดปรกติ

ตัวอย่าง 3.19

พิจารณาข้อมูลที่ประกอบด้วยตัวแปรเชิงตัวเลข 2 ตัวแปร จำนวน 12 ข้อมูล คือ A B C D E F G H I J K และ L แสดงแผนภาพการกระจายข้อมูล ดังภาพ 3.25 และกำหนดค่าพารามิเตอร์ ε เท่ากับรัศมีวงกลมที่ล้อมรอบจุดข้อมูลในภาพ และค่าพารามิเตอร์ $minPts$ เท่ากับ 5

สร้างเซต L ที่มีสมาชิกเป็นข้อมูลทั้งหมด นั่นคือ $L = \{A, B, C, D, E, F, G, H, I, J, K, L\}$ จากแผนภาพการกระจาย พบว่าจุดข้อมูล A G H และ K เป็นจุดแกน ดังนั้น เซต $Core = \{A, G, H, K\}$

ในขณะที่จำนวนกลุ่ม $k = 1$ พิจารณาจุดแกน A สร้างเซต $C_1 = \{A\}$ และหาจุดข้อมูลที่สามารถไปถึงได้ด้วยความหนาแน่นจากจุด A ได้จุด B C D และ E นำจุดข้อมูลเหล่านี้บรรจุเป็นสมาชิกในเซต C_1 จะได้ $C_1 = \{A, B, C, D, E\}$ จากนั้นทำการลบสมาชิกของเซต L และ $Core$ ที่เป็นสมาชิกในเซต C_1 จะได้ $L = \{F, G, H, I, J, K, L\}$ และ $Core = \{G, H, K\}$ ต่อมา นำเซต C_1 เพิ่มเข้าเป็นสมาชิกของเซต C จะได้ $C = \{C_1\}$

ในขณะที่จำนวนกลุ่ม $k = 2$ พิจารณาจุดแกน G สร้างเซต $C_2 = \{G\}$ และหาจุดข้อมูลที่สามารถไปถึงได้ด้วยความหนาแน่นจากจุด G ได้จุด J K I L และ H นำจุดเข้ามูลเหล่านี้เข้าเป็นสมาชิกในเซต C_2 จะได้ $C_2 = \{G, J, K, I, L, H\}$ จากนั้นลบสมาชิกของเซต L และ $Core$ ที่เป็นสมาชิกในเซต C_2 จะได้ $L = \{F\}$ และ $Core = \{\}$ ต่อมา นำเซต C_2 เพิ่มเข้าเป็นสมาชิกของเซต C จะได้ $C = \{C_1, C_2\}$

เมื่อพบว่าเซต $Core$ เป็นเซตว่าง จึงทำการสร้างเซต C_{noise} โดย $C_{noise} = L = \{F\}$ และนำเซต C_{noise} เพิ่มเข้าเป็นสมาชิกของเซต C จะได้ $C = \{C_1, C_2, C_{noise}\}$

จากการจัดกลุ่มข้อมูลทั้ง 12 ข้อมูลด้วยวิธีดีบีสแกน ได้กลุ่มข้อมูลผลลัพธ์ 2 กลุ่ม คือ $C_1 = \{A, B, C, D, E\}$ และ $C_2 = \{G, J, K, I, L, H\}$ ส่วนข้อมูล F จัดเป็นข้อมูลผิดปรกติ

3.3 การวิเคราะห์ความสัมพันธ์

การวิเคราะห์ความสัมพันธ์ (Association Analysis) เป็นการหาความสัมพันธ์ในรูปแบบการเกิดขึ้นร่วมกันหรือกฎความสัมพันธ์ของค่าตัวแปรตั้งแต่ 2 ตัวแปรขึ้นไป มักถูกนำไปใช้ในการวิเคราะห์ข้อมูลรายการเปลี่ยนแปลง (Transaction Data) เช่น ข้อมูลรายการขายสินค้าของร้านค้า ข้อมูลรายการใช้จ่ายผ่านบัตรเครดิต และ ข้อมูลรายการซื้อสินค้าผ่านร้านค้าออนไลน์ เป็นต้น ในที่นี้เรา

จะศึกษาวิธีการวิเคราะห์ความสัมพันธ์ผ่านต้นแบบงานประยุกต์การวิเคราะห์ตะกร้าตลาด (Market Basket Analysis) ซึ่งเป็นการวิเคราะห์ความสัมพันธ์ของสินค้าในตะกร้าสินค้า จากข้อมูลรายการซื้อของลูกค้าทั้งหมด โดยมีเป้าหมาย 2 ข้อ คือ

- ค้นหาชุดของสินค้า เรียกว่า ไอเท็มเซต (Itemset) ที่ถูกซื้อร่วมกัน เช่น การซื้อนมกับขนมปังพร้อมกันในรายการซื้อสินค้าเดียวกัน
- ค้นหากฎความสัมพันธ์ (Association Rule) ระหว่างสินค้าในรายการซื้อของลูกค้า เช่น ถ้ามีการซื้อต้นไม้ประเภทไม้ดอกแล้ว จะมีการปุ๋ยสำหรับบำรุงดอกด้วย

การวิเคราะห์ตะกร้าตลาดนี้ สามารถนำไปใช้ประโยชน์ได้อย่างมาก โดยเฉพาะการศึกษาพฤติกรรมของลูกค้า อันจะนำไปสู่การวางแผนส่งเสริมการขาย การจัดพื้นที่ร้านค้า และการจัดวางสินค้า เป็นต้น

ไอเท็มเซต

ให้ $\mathscr{I} = \{X_1, X_2, ...X_d\}$ เป็นชุดของสิ่งสนใจทั้งหมด เซตย่อย (Subset) หนึ่งๆ ของเซต $\mathscr{I}$ ยกเว้นเซตว่าง จะถูกเรียกว่า ไอเท็มเซต ในกรณีการวิเคราะห์ตะกร้าตลาด เซต $\mathscr{I}$ คือ สินค้าทั้งหมดที่ลูกค้าสามารถหยิบใส่ตะกร้าสินค้าได้ และไอเท็มเซต คือ ชุดของสินค้าในตะกร้าสินค้าหนึ่งๆ ตัวอย่างเช่น ร้านผลไม้แห่งหนึ่งมีผลไม้ขายภายในร้าน 5 ชนิด คือ ส้ม แอปเปิล แตงโม มะละกอ และกล้วย ดังนั้น เซต $\mathscr{I}$ = {ส้ม, แอปเปิล, แตงโม, มะละกอ, กล้วย} ซึ่งเป็นสินค้าที่ลูกค้าสามารถหยิบใส่ตะกร้าสินค้าได้ ลูกค้าคนแรกเลือกหยิบ ส้มและแอปเปิล และลูกค้าอีกคนหนึ่งเลือกหยิบ แตงโม มะละกอ และส้ม ทั้งเซต {ส้ม, แอปเปิล} และ {แตงโม, มะละกอ, ส้ม} ต่างก็เป็นไอเท็มเซตที่เกิดขึ้นภายในร้านผลไม้นี้ เมื่อเซต $\mathscr{I}$ มีจำนวนสมาชิกเท่ากับ m ตัว จะได้ว่าไอเท็มเซตที่เป็นไปได้ทั้งหมดมีจำนวน $2^m - 1$ ไอเท็มเซต นั่นคือ เซตย่อยทั้งหมดของเซต $\mathscr{I}$ ไม่รวมเซตว่าง

ตัวอย่าง 3.20

ร้านขายอาหารเช้าริมทางร้านหนึ่ง มีสินค้าขายทั้งหมด 4 ชนิด คือ กล้วย นม แอปเปิล และขนมปัง ดังนั้น ไอเท็มเซตของสินค้าที่ลูกค้าสามารถซื้อพร้อมกันได้สำหรับร้านค้าแห่งนี้ มีทั้งหมด $2^4 - 1 = 15$ ไอเท็มเซต ที่เป็นไปได้ ดังนี้

{กล้วย} {นม} {แอปเปิล} {ขนมปัง}
{กล้วย, นม} {กล้วย, แอปเปิล} {กล้วย, ขนมปัง} {นม, แอปเปิล}
{นม, ขนมปัง} {แอปเปิล, ขนมปัง}
{กล้วย, นม, แอปเปิล} {กล้วย, นม, ขนมปัง}

{กล้วย, แอปเปิล, ขนมปัง} {นม, แอปเปิล, ขนมปัง}
และ {กล้วย, นม, แอปเปิล, ขนมปัง}

กำหนดให้ชุดข้อมูล D จัดเก็บข้อมูลรายการเปลี่ยนแปลงในรูปแบบฐานข้อมูลทวิภาค (Binary Database) ที่ประกอบด้วยคอลัมน์จำนวน d คอลัมน์ แต่ละคอลัมน์แทนสินค้าหนึ่งๆ และแถวข้อมูลจำนวน n แถว โดยแต่ละแถวแทนรายการซื้อ/ขาย รายการที่ i ประกอบด้วยสินค้า X_j ก็ต่อเมื่อ ฐานข้อมูลทวิภาคแถวที่ i คอลัมน์ที่ j มีค่าเป็น 1 ตาราง 3.1 แสดงตัวอย่างข้อมูลรายการเปลี่ยนแปลงในรูปแบบฐานข้อมูลทวิภาค ที่ประกอบด้วยสินค้าทั้งหมด 4 สินค้า และรายการซื้อ 10 รายการ รายการ 5 แสดงการซื้อสินค้า 3 สินค้า คือ นม แอปเปิล และ ขนมปัง ในหัวข้อนี้เราจะใช้การแทนข้อมูลรายการเปลี่ยนแปลงในรูปแบบฐานข้อมูลทวิภาคในการอธิบายรายละเอียดของการวิเคราะห์ตะกร้าตลาด

	Banana	Milk	Apple	Bread
t_1	0	1	1	0
t_2	1	1	0	0
t_3	0	1	0	1
t_4	1	0	1	0
t_5	0	1	1	1
t_6	1	1	0	1
t_7	0	1	1	1
t_8	0	0	1	0
t_9	0	1	0	1
t_{10}	1	0	1	1

ตาราง 3.1: ข้อมูลรายการซื้อสินค้าจากร้านขาย อาหาร เช้า ริม ทาง แห่ง หนึ่ง ซึ่งมี สินค้าวางขาย 4 ชนิด คือ กล้วย (Banana) นม (Milk) แอปเปิล (Apple) และขนมปัง (Bread) จำนวน 10 รายการซื้อ เมื่อค่า 1 แทน สถานะการมีสินค้าชนิดนั้นในรายการซื้อ และค่า 0 แทน สถานะการไม่มีสินค้าชนิดนั้นในรายการซื้อ

ค่าสนับสนุน (Support) ของไอเท็มเซตหนึ่งๆ เป็นค่าสัดส่วนของจำนวนรายการในชุดข้อมูลที่ไอเท็มเซตเป็นส่วนหนึ่งของสมาชิก ต่อจำนวนรายการทั้งหมด ค่าสนับสนุนของไอเท็มเซต $I = \{X_p, ..., X_q\}$ ในชุดข้อมูล D แทนด้วยสัญลักษณ์ $sup(I, \mathrm{D})$ สามารถคำนวณได้ด้วยสมการต่อไปนี้

$$sup(I, \mathrm{D}) = \frac{|\{t|t \in \mathrm{D} \text{ and } I \subseteq t|\}}{n} \tag{3.32}$$

เมื่อ n คือ จำนวนรายการในชุดข้อมูล D

เราจะกล่าวว่า ไอเท็มเซต I เป็นไอเท็มเซตถี่ (Frequent Itemset) ในชุดข้อมูล D ถ้า $sup(I, \mathrm{D}) \geq minsup$ เมื่อ $minsup$ คือ ค่าขีดแบ่งค่าสนับสนุนต่ำสุดซึ่งจะต้องถูกกำหนดขึ้น

ตัวอย่าง 3.21

จากชุดข้อมูลรายการซื้อสินค้าในตาราง 3.1 เราสามารถหาค่าสนับสนุนของแต่ละไอเท็มเซตที่เป็นไปได้ ได้ดังนี้

ค่าสนับสนุน	ไอเท็มเซต
0.7	{นม}
0.6	{แอปเปิล} และ{ขนมปัง}
0.5	{นม, ขนมปัง}
0.4	{กล้วย}
0.3	{นม, แอปเปิล} และ {แอปเปิล, ขนมปัง}
0.2	{กล้วย, นม} และ {กล้วย, แอปเปิล} และ {กล้วย, ขนมปัง} และ {นม, แอปเปิล, ขนมปัง}
0.1	{กล้วย, นม, ขนมปัง} และ {กล้วย, แอปเปิล, ขนมปัง}

เมื่อกำหนดค่า $minsup = 0.5$ จะได้ว่าไอเท็มเซตต่อไปนี้เป็นไอเท็มเซตถี่

{นม} {แอปเปิล} {ขนมปัง} และ {นม, ขนมปัง}

เนื่องจากมีค่าสนับสนุนมากกว่าหรือเท่ากับค่าขีดแบ่งค่าสนับสนุนต่ำสุดที่กำหนด

การค้นหาไอเท็มเซตถี่ในชุดข้อมูล D ที่มีเซตของสินค้าที่สนใจ $\mathscr{I} = \{X_1, X_2, ..., X_d\}$ และกำหนดค่าขีดแบ่งค่าสนับสนุนต่ำสุด $minsup$ สามารถดำเนินการได้อย่างอัตโนมัติ โดยการใช้ขั้นตอนวิธีเอไพรออริ (Apriori Algorithm) ซึ่งมีขั้นตอน ดังนี้

ขั้นตอนที่ 1 สร้างชุดของไอเท็มเซต C_1 ที่มีสมาชิกเพียง 1 สินค้า โดย $C_1 = \{\{X_1\}, \{X_2\}, ..., \{X_d\}\}$

ขั้นตอนที่ 2 คำนวณค่าสนับสนุนของแต่ละไอเท็มเซตในเซต C_1 บนชุดข้อมูล D

ขั้นตอนที่ 3 สร้างเซต L_1 โดยคัดเลือกไอเท็มเซตในเซต C_1 ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับ $minsup$ จะได้ $L_1 = \{I | I \in C_1 \textbf{ and } sup(I, \mathrm{D}) \geq minsup\}$

ขั้นตอนที่ 4 กำหนดค่า $k = 2$

ขั้นตอนที่ 5 ทำขั้นตอนต่อไปนี้ จนกระทั่งเซต L_{k-1} เป็นเซตว่าง

5.1 สร้างชุดของไอเท็มเซต C_k โดยการเชื่อม (Join) เซต L_{k-1} และ L_{k-1} เข้าด้วยกัน โดยสมาชิกในเซต L_{k-1} จะนำมาเชื่อมกันได้ ถ้ามีสินค้าร่วมกันจำนวน $k - 2$ สินค้า

5.2 คำนวณค่าสนับสนุนของแต่ละไอเท็มเซตในเซต C_k บนชุดข้อมูล D

5.3 สร้างเซต L_k โดยคัดเลือกไอเท็มเซตในเซต C_k ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับ $minsup$ จะได้ $L_k = \{I | I \in C_k \textbf{ and } sup(I, \mathrm{D}) \geq minsup\}$

5.4 เพิ่มค่า k ขึ้น 1

ขั้นตอนที่ 6 รวมเซต L_i สำหรับ $i = 1, 2, ..., k - 2$ เป็นเซตของไอเท็มเซตถี่

จากขั้นตอนวิธีเอไพรออริ เราจะเห็นว่าเป็นสร้างและค้นหาไอเท็มเซตถี่ โดยเริ่มสร้างไอเท็มเซตที่มีประกอบด้วยสินค้า 1 สินค้า ในขั้นตอนที่ 1 แล้วค่อยๆ สร้างไอเท็มเซตที่มีขนาดเพิ่มขึ้นเรื่อยๆ โดยอาศัยไอเท็มเซตที่มีจำนวนสินค้าน้อยกว่า 1 สินค้า ในขั้นตอนที่ 5 (กรณีค่า $k \geq 2$) เป็นการสร้างไอเท็มเซตที่ประกอบด้วยสินค้ามากกว่าหรือเท่ากับ 2 สินค้า (ในขั้นตอนที่ 5.1) โดยอาศัยไอเท็มเซตถี่ที่ประกอบด้วยจำนวนสินค้าน้อยกว่า 1 สินค้า โดยการจับคู่ไอเท็มเซตเหล่านั้นที่มีสินค้าเหมือนกัน $k - 2$ สินค้า เพื่อให้ได้ไอเท็มเซตที่มีสินค้า k ชิ้น ตัวอย่างเช่น การสร้างไอเท็มเซตที่ประกอบด้วยสินค้า 5 ชิ้น จากไอเท็มเซตที่ประกอบด้วยสินค้า 4 ชิ้นที่มีสมาชิกเหมือนกัน 3 สินค้า เนื่องจากสมาชิกในเซตจะต้องมีค่าไม่เหมือนกัน การรวมเซต 2 เซต สมาชิกเหมือนกันจะถูกนำเข้าเซตเพียงแค่ 1 ครั้ง เช่น การรวมเซต $\{A, B, C, D\}$ และ $\{A, B, C, E\}$ จะได้เซตผลลัพธ์ $\{A, B, C, D, E\}$ ดังนั้น การรวมไอเท็มเซตที่ประกอบด้วยสินค้า 4 ชิ้นที่มีสมาชิกเหมือนกัน 3 สินค้า ได้ผลลัพธ์ไอเท็มเซตที่ประกอบด้วยสินค้า 5 สินค้า นอกจากนี้เมื่อพิจารณาขั้นตอนที่ 3 และ 5.3 แล้ว จะพบว่ามีการตัดไอเท็มเซตที่มีค่าสนับสนุนน้อยกว่าค่า $minsup$ ซึ่งเป็นการคัดกรองไอเท็มเซตถี่ โดยผลลัพธ์จะถูกนำไปสร้างไอเท็มเซตใหม่ที่ประกอบด้วยสินค้าเพิ่มขึ้น 1 สินค้า (ในขั้นตอนที่ 5.1) การพิจารณาเฉพาะไอเท็มเซตถี่ทำให้เวลาในการดำเนินการลดลง เนื่องจากไอเท็มเซตที่ไม่ได้เกิดขึ้นบ่อยครั้งจะไม่สามารถสร้างไอเท็มเซตถี่ใหม่ที่ประกอบด้วยสินค้าเพิ่มขึ้น 1 สินค้าได้

ตัวอย่าง 3.22

จากชุดข้อมูลรายการซื้อสินค้าในตาราง 3.1 สามารถค้นหาไอเท็มเซตถี่โดยใช้ขั้นตอนวิธีเอไพรออริในการ เมื่อกำหนดค่า $minsup = 0.5$ ได้ดังนี้

สร้างชุดไอเท็มเซตที่ประกอบด้วยสินค้า 1 สินค้า และคำนวณค่าสนับสนุนของแต่ละไอเท็มเซต จะได้

C_1	ค่าสนับสนุน
{กล้วย}	0.4
{นม}	0.7
{แอปเปิล}	0.6
{ขนมปัง}	0.6

ต่อมาสร้างเซต L_1 บรรจุไอเท็มเซตใน C_1 ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่า $minsup$ จะได้

$$L_1 = \{\{\text{นม}\}, \{\text{แอปเปิล}\}, \{\text{ขนมปัง}\}\}$$

ในขณะที่ค่า $k = 2$ ทำการสร้างชุดของไอเท็มเซต C_2 จากการเชื่อมเซต L_1 และ L_1 โดยที่ไอเท็มเซตที่สามารถเชื่อมกันได้ จะต้องมีสมาชิกร่วมกัน $k - 2 = 2 - 2 = 0$ สมาชิก พร้อมทั้งหาค่าสนับสนุนของไอเท็มเซตผลลัพธ์ จะได้

C_2	ค่าสนับสนุน
{กล้วย, นม}	0.2
{กล้วย, แอปเปิล}	0.2
{กล้วย, ขนมปัง}	0.2
{นม, แอปเปิล}	0.3
{นม, ขนมปัง}	0.5
{แอปเปิล, ขนมปัง}	0.3

ต่อมาสร้างเซต L_2 บรรจุไอเท็มเซตใน C_2 ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่า $minsup$ จะได้

$$L_2 = \{\{\text{นม}, \text{ขนมปัง}\}\}$$

ในขณะที่ค่า $k = 3$ ทำการสร้างชุดของไอเท็มเซต C_3 จากการเชื่อมเซต L_2 และ L_2 โดยที่ไอเท็มเซตที่สามารถเชื่อมกันได้ จะต้องมีสมาชิกร่วมกัน $k - 2 = 3 - 2 = 1$ สมาชิก พร้อมทั้งหาค่าสนับสนุนของไอเท็มเซตผลลัพธ์ พบว่าไม่สามารถนำไอเท็มเซตใดใน L_2 มาสร้างเป็นไอเท็มเซตใหม่ได้ จึงทำให้เซต C_3 เป็นเซตว่าง และทำให้เซต L_3 เป็นเซตว่างตามไปด้วย

ในขณะที่ค่า $k = 4$ เมื่อเซต L_3 เป็นเซตว่าง จึงหยุดกระบวนการค้นหาไอเท็มเซตถี่

ดังนั้น ไอเท็มเซตถี่ที่ได้จากกระบวนการค้นหาด้วยวิธีเอไพรออริ คือสมาชิกของ $L_1 \cup L_2$ นั่นคือ

{นม} {แอปเปิล} {ขนมปัง} และ {นม, ขนมปัง}

กฎความสัมพันธ์

กฎความสัมพันธ์เป็นการอธิบายความสัมพันธ์ระหว่างไอเท็มเซต 2 เซตในลักษณะการเกิดขึ้นร่วมกันอย่างมีเงื่อนไข โดยอธิบายในรูปแบบกฎถ้า-แล้ว (If-then Rule) ตัวอย่างเช่น ถ้าลูกค้าซื้อนมและแอปเปิลร่วมกันแล้ว เขามีแนว

โน้มจะซื้อกล้วยด้วย กำหนดให้ I และ J เป็นไอเท็มเซต โดยที่ I และ J ไม่มีสมาชิกร่วมกัน และ $I, J \subseteq \mathscr{I}$ กฎความสัมพันธ์ $I \Rightarrow J$ แสดงความสัมพันธ์ระหว่างไอเท็มเซต I และ J ว่า ถ้าลูกค้าซื้อสินค้าในไอเท็มเซต I ร่วมกันแล้ว เขามีแนวโน้มจะซื้อสินค้าทั้งหมดในไอเท็มเซต J ด้วย เราจะเรียก I ว่า ส่วนซ้ายมือ (Left-Hand-Side: LHS) และเรียก J ว่า ส่วนขวามือ (Right-Hand Side: RHS) และพึงทราบประการหนึ่งว่า กฎความสัมพันธ์ $I \Rightarrow J$ ไม่เท่ากับกฎความสัมพันธ์ $J \Rightarrow I$ เนื่องจากการดำเนินการ ถ้า-แล้ว ไม่มีคุณสมบัติการสลับที่ (Commutative Property) ตัวอย่างเช่น จากกฎความสัมพันธ์ที่ว่า ถ้าลูกค้าซื้อนมและแอปเปิลร่วมกันแล้ว เขามีแนวโน้มจะซื้อกล้วยด้วย เราจะสามารถเขียนในรูปสัญลักษณ์ได้ ดังนี้ {นม, แอปเปิล} $\Rightarrow$ {กล้วย} โดยที่ ไอเท็มเซต {นม, แอปเปิล} เป็นส่วนซ้ายมือของกฎ และไอเท็มเซต {กล้วย} เป็นส่วนขวามือของกฎ

ค่าสนับสนุน ของกฎความสัมพันธ์ คือ ค่าสัดส่วนของจำนวนรายการในชุดข้อมูลที่ไอเท็มเซต $I \cup J$ เป็นส่วนหนึ่งของสมาชิก ต่อจำนวนรายการทั้งหมด กล่าวอีกนัยหนึ่งคือค่าสนับสนุนของไอเท็มเซต $I \cup J$ นั่นเอง สามารถเขียนในรูปสมการคณิตศาสตร์ ได้ดังนี้

$$sup(I \Rightarrow J) = sup(I \cup J) \tag{3.33}$$

เมื่อ $sup(I \Rightarrow J)$ คือ ค่าสนับสนุนของกฎความสัมพันธ์ $I \Rightarrow J$

ค่าความเชื่อมั่น (Confidence) เป็นค่าที่แสดงว่ากฎความสัมพันธ์หนึ่งๆ เป็นจริงบ่อยครั้งเพียงใดในชุดข้อมูล D สามารถคำนวณได้ ดังนี้

$$conf(I \Rightarrow J) = \frac{sup(I \cup J)}{sup(I)} \tag{3.34}$$

เมื่อ $conf(I \Rightarrow J)$ คือ ค่าความเชื่อมั่นของกฎความสัมพันธ์ $I \Rightarrow J$

เราจะกล่าวว่า กฎความสัมพันธ์ $I \Rightarrow J$ เป็นกฎที่เกิดขึ้นบ่อยครั้งในชุดข้อมูล D ถ้า $sup(I \Rightarrow J) \geq minsup$ และจะกล่าวว่า เป็นกฎที่น่าเชื่อถือในชุดข้อมูล D ถ้า $conf(I \Rightarrow J) \geq minconf$ เมื่อ $minsup$ และ $minconf$ คือ ค่าขีดแบ่งค่าสนับสนุนต่ำสุด และค่าขีดแบ่งค่าความเชื่อมั่นต่ำสุด ที่จะต้องถูกกำหนดขึ้น

ค่าลิฟท์ (Lift) เป็นค่าอัตราส่วนระหว่างค่าสนับสนุนของกฎความสัมพันธ์และค่าสนับสนุนคาดหวัง ซึ่งแสดงถึงผลกระทบที่เกิดจากกฎความสัมพันธ์ สามารถคำนวณได้จากสูตรต่อไปนี้

$$lift(I \Rightarrow J) = \frac{sup(I \cup J)}{sup(I) \times sup(J)} \tag{3.35}$$

ค่าลิฟท์ของกฎความสัมพันธ์ $I \Rightarrow J$ เขียนแทนด้วยสัญลักษณ์ $lift(I \Rightarrow J)$ มีค่าอยู่ระหว่าง 0 และค่าอนันต์ ถ้าค่าลิฟท์มีค่าเท่ากับ 1 บ่งบอกว่าความน่าจะเป็นของการเกิดไอเท็มเซตในส่วนซ้ายมือในตะกร้าสินค้าและไอเท็มเซตส่วนขวามือเป็นอิสระต่อกัน ค่าลิฟท์ที่มีค่ามากกว่า 1 แสดงถึงระดับความขึ้นต่อกันของความน่าจะเป็นของการเกิดไอเท็มเซตในส่วนซ้ายมือและไอเท็มเซตส่วนขวามือในชุดข้อมูล การมีไอเท็มเซตส่วนซ้ายมือในตระกร้าสินค้าส่งผลให้ความน่าจะเป็นของการเกิดไอเท็มเซตทางขวามือในตะกร้าสินค้าเพิ่มขึ้นด้วย และถ้าค่าลิฟท์มีค่าน้อยกว่า 1 แสดงนัยว่าไอเท็มเซตส่วนซ้ายมือในตระกร้าสินค้าส่งผลให้ความน่าจะเป็นของการเกิดไอเท็มเซตทางขวามือในตะกร้าสินค้าลดลง

ตัวอย่าง 3.23

พิจารณากฎความสัมพันธ์ {นม} $\Rightarrow$ {ขนมปัง} จากข้อมูลรายการซื้อสินค้าในตาราง 3.1 สามารถหาค่าสนับสนุนของกฎความสัมพันธ์นี้ ได้ดังนี้

$$\begin{aligned} sup(\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}) &= sup(\{\text{นม}\} \cup \{\text{ขนมปัง}\}) \\ &= sup(\{\text{นม}, \text{ขนมปัง}\}) \\ &= 0.5 \end{aligned}$$

และความเชื่อมั่นของกฎความสัมพันธ์ สามารถคำนวณ ดังนี้

$$\begin{aligned} conf(\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}) &= \frac{sup(\{\text{นม}, \text{ขนมปัง}\})}{sup(\{\text{นม}\})} \\ &= \frac{0.5}{0.7} = 0.71 \end{aligned}$$

เมื่อกำหนดค่า $minsup = 0.5$ และ $minconf = 0.3$ พบว่า กฎความสัมพันธ์ {นม} $\Rightarrow$ {ขนมปัง} มีค่าสนับสนุนมากกว่าค่า $minsup$ ดังนั้น กฎความสัมพันธ์นี้จึงเป็นกฎที่เกิดขึ้นบ่อยครั้ง นอกจากนี้ยังพบว่าค่าความเชื่อมั่นของกฎความสัมพันธ์มีค่าสนับสนุนมากกว่าค่า $minconf$ ดังนั้น กฎความสัมพันธ์นี้จึงเป็นกฎน่าเชื่อถือด้วย

นอกจากนี้ เรายังสามารถหาค่าลิฟท์ของกฎความสัมพันธ์ ได้ดังนี้

$$\begin{aligned} lift(\{\text{นม}\} \Rightarrow \{\text{ขนมปัง}\}) &= \frac{sup(\{\text{นม}\} \cup \{\text{ขนมปัง}\})}{sup(\{\text{นม}\}) \times sup(\{\text{นม}\})} \\ &= \frac{sup(\{\text{นม}, \text{ขนมปัง}\})}{sup(\{\text{นม}\}) \times sup(\{\text{นม}\})} \\ &= \frac{0.5}{0.7 \times 0.6} \\ &= \frac{0.5}{0.49} = 1.19 \end{aligned} \tag{3.36}$$

จากค่าลิฟท์ของกฎความสัมพันธ์ {นม} $\Rightarrow$ {ขนมปัง} มีค่าเท่ากับ 1.19 ซึ่งมีค่ามากกว่า 1 ดังนั้น เราจึงสามารถกล่าวได้ว่าถ้ามีนมอยู่ในตะกร้าสินค้าของลูกค้าแล้ว มีแนวโน้มสูงที่ในตะกร้าสินค้านั้นจะมีขนมปังด้วย

เราสามารถค้นหากฎความสัมพันธ์ที่มีความน่าเชื่อถือจากชุดของไอเท็มเซตถี่ $\mathscr{F}$ ในชุดข้อมูล D ได้อย่างอัตโนมัติ โดยการกำหนดค่า $minconf$ และดำเนินการตามขั้นตอนวิธี ต่อไปนี้

สำหรับแต่ละไอเท็มเซต Z ในเซต $\mathscr{F}$ ที่มีจำนวนสินค้าในไอเท็มเซตมากกว่า 2 สินค้า ดำเนินการขั้นตอนดังนี้

ขั้นตอนที่ 1 สร้างเซต $\mathscr{A}$ ซึ่งเป็นเซตของเซตย่อยทั้งหมดของ I ยกเว้นเซตว่าง โดย $\mathscr{A} = \{I | I \subset Z, I \neq \emptyset\}$

ขั้นตอนที่ 2 ทำขั้นตอนต่อไปนี้จนกระทั่ง เซต $\mathscr{A}$ เป็นเซตว่าง

2.1 ให้ไอเท็มเซต I เป็นไอเท็มเซตใน $\mathscr{A}$ ที่มีขนาดใหญ่ที่สุด (จำนวนสมาชิกมากที่สุด)

2.2 นำไอเท็มเซต I ออกจากการเป็นสมาชิกของเซต $\mathscr{A}$

2.3 สร้างไอเท็มเซต J โดย J คือ ผลลัพธ์จากการนำสินค้าที่อยู่ในไอเท็มเซต I ออกจากไอเท็มเซต Z นั่นคือ $J = Z \setminus I$

2.4 คำนวณค่าความเชื่อมั่น c โดย $c = \dfrac{sup(Z)}{sup(I)}$

2.5 ถ้าค่า c มีค่ามากกว่าหรือเท่ากับ $minconf$ แล้ว สร้างกฎความสัมพันธ์ $I \Rightarrow J$ ถ้าไม่เช่นนั้น ทำให้การลบไอเท็มเซตที่เป็นเซตย่อยทั้งหมดของ I ออกจากเซต $\mathscr{A}$

จากขั้นตอนวิธีสำหรับค้นหากฎความสัมพันธ์ที่น่าเชื่อถือข้างต้น เป็นการนำแต่ละไอเท็มเซตถี่ Z ที่ประกอบด้วยจำนวนสินค้ามากกว่า 2 สินค้าขึ้นไปทั้งหมดมาพิจารณา โดยการสร้างไอเท็มเซตส่วนซ้ายมือของกฎทั้งหมดที่เป็นไปได้จากไอเท็มเซตถี่ Z ในขั้นตอนที่ 1 ต่อมานำแต่ละไอเท็มเซตส่วนซ้ายมือ I มาพิจารณาสร้างส่วนขวามือ J โดยส่วนขวามือเป็นไอเท็มเซตของสินค้าที่เหลือจากการนำสินค้าในส่วนซ้ายมือออกจากไอเท็มเซตถี่ Z และทำการคำนวณค่าความเชื่อมั่นที่จะเกิดจากการสร้างกฎความสัมพันธ์ $I \Rightarrow J$ ถ้าค่าความเชื่อมั่นที่คำนวณมีค่ามากกว่าหรือเท่ากับค่า $minconf$ จะทำการสร้างกฎความเชื่อมั่น $I \Rightarrow J$ แต่ถ้าค่าความเชื่อมั่นที่คำนวณมีค่าน้อยกว่าค่า $minconf$ จะทำการลบเซตย่อยทั้งหมดที่เป็นไปได้ของไอเท็มเซต I ออกจากเซตของไอเท็มเซตที่มีศักยภาพที่จะเป็นส่วนซ้ายมือของกฎก่อนการนำไอเท็มเซตที่มีศักยภาพที่จะเป็นส่วนซ้ายมืออื่นขึ้นมาพิจารณา

ตัวอย่าง 3.24

จากตัวอย่าง 3.22 ไอเท็มเซตถี่ในชุดข้อมูลรายการสินค้าในตาราง 3.1 มี

จำนวน 4 ไอเท็มเซต คือ {นม} {แอปเปิล} {ขนมปัง} และ {นม, ขนมปัง} พบว่ามีเพียง 1 ไอเท็มเซตถี่ที่มีจำนวนสินค้ามากกว่า 2 สินค้า นั่นคือ Z = {นม, ขนมปัง} เมื่อกำหนดค่า $minconf = 0.3$ เราสามารถสร้างกฎความสัมพันธ์ที่น่าเชื่อถือจากไอเท็มเซตถี่ดังกล่าว ได้ดังนี้

สร้างเซต $\mathscr{A}$ ซึ่งเป็นเซตของเซตย่อยทั้งหมดของ {นม, ขนมปัง} จะได้ $\mathscr{A}$ = {{นม}, {ขนมปัง}}

พิจารณาให้ไอเท็มเซต I = {นม} ซึ่งมีขนาดใหญ่ที่สุด เป็นส่วนซ้ายมือของกฎความสัมพันธ์ และนำเซต I ออกจากเซต $\mathscr{A}$ จะได้ $\mathscr{A}$ = {ขนมปัง}} ต่อมาสร้างไอเท็มเซต $J = Z \setminus I$ = {นม, ขนมปัง} – {นม} = {ขนมปัง} จากนั้นคำนวณค่าความเชื่อมั่น $c = \frac{\{\text{นม, ขนมปัง}\}}{\{\text{นม}\}} = \frac{0.5}{0.7} = 0.71$ เนื่องจากค่าความเชื่อมั่นที่คำนวณได้มีค่ามากกว่าค่า $minconf$ ดังนั้น จึงสร้างกฎความสัมพันธ์ {นม} ⟹ {ขนมปัง}

พิจารณาให้ไอเท็มเซต I = {ขนมปัง} เป็นส่วนซ้ายมือของกฎความสัมพันธ์ และนำเซต I ออกจากเซต $\mathscr{A}$ จะได้ $\mathscr{A}$ = {} ต่อมาสร้างไอเท็มเซต $J = Z \setminus I$ = {นม, ขนมปัง} – {ขนมปัง} = {นม} จากนั้นคำนวณค่าความเชื่อมั่น $c = \frac{\{\text{นม, ขนมปัง}\}}{\{\text{ขนมปัง}\}} = \frac{0.5}{0.6} = 0.83$ เนื่องจากค่าความเชื่อมั่นที่คำนวณได้มีค่ามากกว่าค่า $minconf$ ดังนั้น จึงสร้างกฎความสัมพันธ์ {ขนมปัง} ⟹ {นม}

เมื่อพบว่าเซต $\mathscr{A}$ เป็นเซตว่าง จึงหยุดการสร้างกฎความสัมพันธ์ที่น่าเชื่อถือจากไอเท็มเซตถี่ Z อีกทั้งไม่มีไอเท็มเซตถี่อื่นที่มีจำนวนสินค้ามากกว่า 2 สินค้าที่จะสามารถนำมาสร้างกฎความสัมพันธ์ได้อีก ดังนั้น จากชุดข้อมูลรายการสินค้านี้ มีกฎความสัมพันธ์ที่น่าเชื่อถือ 2 กฎ คือ

{นม} ⟹ {ขนมปัง} มีค่าความเชื่อมั่นเท่ากับ 0.71

{ขนมปัง} ⟹ {นม} มีค่าความเชื่อมั่นเท่ากับ 0.83

3.4 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. Bishop, Christopher M. (2006). ***Pattern Recognition and Machine Learning (Information Science and Statistics)***. Berlin, Heidelberg: Springer-Verlag.
2. Joanes, D. N. and C. A. Gill (1998). "Comparing measures of sample skewness and kurtosis." In: ***Journal of the Royal Statistical Society: Series D (The Statistician)*** 47.1, pp. 183–189.
3. Raymond H. Meyers, Ronald E. Walpole and, Sharon L. Meyers, and Keying E. Ye (2012). ***Probability & Statistics for Engineers & Scientists***. 9th. USA: Prentice Hall.
4. Theodoridis, Sergios and Konstantinos Koutroumbas (2008). ***Pattern Recognition***. 4th. USA: Academic Press, Inc.
5. Westfall, Peter H. (2014). "Kurtosis as Peakedness" In: ***The American Statistician*** 68.3, pp. 191–195.
6. Zaki, Mohammed J. and Wagner Meira Jr (2014). ***Data Mining and Analysis: Fundamental Concepts and Algorithms***. USA: Cambridge University Press.

3.5 แบบฝึกหัดท้ายบท

1. จงหาค่ากลางข้อมูล (ค่าเฉลี่ย ค่ามัธยมฐาน และค่าฐานนิยม) ส่วนเบี่ยงเบนมาตรฐาน และความแปรปรวน ของข้อมูลในชุดข้อมูลต่อไปนี้ (หมายเหตุ บางชุดข้อมูลอาจไม่สามารถคำนวณค่าสถิติเชิงพรรณนาบางค่าได้)
   a) ข้อมูลระยะเวลา (หน่วย: ชั่วโมง) ของการแห้งของสีน้ำพลาสติกยี่ห้อหนึ่ง จากการทดลอง 15 ครั้ง มีข้อมูลดังนี้

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

   b) ข้อมูลค่าพีเอช (pH) ของสารที่มีความเป็นกลาง (pH=7) ที่ตรวจวัดได้จากเครื่องมือวัดเครื่องหนึ่ง โดยทำการวัดซ้ำ 10 ครั้ง มีข้อมูล ดังนี้

7.07	7.00	7.10	6.97	7.00	7.03	7.01	7.01	6.98	7.08

   c) ข้อมูลระดับความพึ่งพอใจในการให้บริการจัดส่งสินค้าของบริษัทแห่งหนึ่ง ระดับความพึ่งพอใจมี 5 ระดับ ได้แก่ ไม่พอใจ พอใจน้อย พอใจปานกลาง พอใจมาก และพอใจมากที่สุด จากผลการสำรวจผู้ใช้บริการ 15 คน มีข้อมูล ดังนี้

พอใจปานกลาง	พอใจมาก	พอใจมาก	พอใจน้อย	พอใจมากที่สุด
ไม่พอใจ	พอใจมาก	พอใจมากที่สุด	พอใจน้อย	พอใจมาก
พอใจปานกลาง	พอใจมาก	พอใจปานกลาง	พอใจน้อย	พอใจมากที่สุด

d) ข้อมูลสำรวจสีขนของแมวที่มีขนสีเดียว จำนวน 10 ตัว ใบบริเวณมหาวิทยาลัยเชียงใหม่ มีข้อมูล ดังนี้

ขาว ขาว ดำ แดง เทา ดำ ดำ ดำ แดง ดำ

2. จากชุดข้อมูล Iris สามารถคำนวณค่าสถิติเชิงพรรณาสำหรับแต่ละตัวแปร ดังนี้

ค่าสถิติ	ตัวแปร			
	Sepal Length	Sepal Width	Petal Length	Petal Width
Mean	5.84	3.05	3.76	1.12
Median	5.80	3.00	4.35	1.30
Mode	5.00	3.00	1.50	0.20
SD	0.83	0.43	1.76	0.76
Skewness	0.31	0.33	-0.27	-0.10
Kurtosis	-0.55	0.29	-1.40	-1.34

จงวาดลักษณะการกระจายของแต่ละตัวแปร พร้อมระบุค่าเฉลี่ย ค่ามัธยมฐาน ค่าฐานนิยม และลักษณะการแจกแจงของค่าตัวแปร

3. จากชุดข้อมูล Iris สามารถคำนวณค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรแต่ละคู่ และแสดงด้วยเมทริกซ์สหสัมพันธ์เพียร์สัน ดังนี้

$$\begin{array}{r c}
 & \begin{array}{cccc} \text{Sepal Length} & \text{Sepal Width} & \text{Petal Length} & \text{Petal Width} \end{array} \\
\begin{array}{r} \text{Sepal Length} \\ \text{Sepal Width} \\ \text{Petal Length} \\ \text{Petal Width} \end{array} & \begin{bmatrix} 1.00 & -0.11 & 0.87 & 0.82 \\ -0.11 & 1.00 & -0.42 & -0.36 \\ 0.87 & -0.42 & 1.00 & 0.96 \\ 0.82 & -0.36 & 0.96 & 1.00 \end{bmatrix}
\end{array}$$

จากเมทริกซ์สหสัมพันธ์เพียร์สันข้างต้น จงตอบคำถาม ต่อไปนี้

a) ตัวแปรคู่ใดมีความสัมพันธ์ระหว่างค่าตัวแปรมากที่สุด
b) ตัวแปรคู่ใดมีความสัมพันธ์ระหว่างค่าตัวแปรน้อยที่สุด
c) จงอธิบายความสัมพันธ์ระหว่างตัวแปร Sepal Length และ Petal Length
d) จงอธิบายความสัมพันธ์ระหว่างตัวแปร Sepal Width และ Petal Width

4. กำหนดข้อมูล 5 ข้อมูล ต่อไปนี้

$$x_1 = (1.00, 2.45, 3.05) \quad x_2 = (2.20, 1.05, 4.00) \quad x_3 = (1.50, 0.05, 2.75)$$
$$x_4 = (5.60, 3.45, 1.95) \quad x_5 = (3.30, 3.00, 3.50)$$

จงสร้างเมทริกซ์ระยะห่างของข้อมูลทั้ง 5 ข้อมูล โดยใช้วิธีการวัดระยะทางแบบยุคลิด การวัดระยะทางแบบแมนฮัตตัน และการวัดความคล้ายแบบโคไซน์

5. จงคำนวณระยะห่างแบบแฮมมิงและค่าสัมประสิทธิ์ความคล้ายแจ็คการ์ดระหว่างข้อมูลนักศึกษามหาวิทยาลัยเชียงใหม่ 2 คน ต่อไปนี้

s_1 = (Science, Computer Science, Male, Frist year)

s_2 = (Science, Mathematics, Male, Male, Second year)

6. จากชุดข้อมูล ต่อไปนี้

	X_1	X_2
x_1	4.03	2.25
x_2	3.73	0.64
x_3	2.70	1.23
x_4	3.29	1.15
x_5	2.21	-0.16
x_6	11.44	5.97
x_7	10.33	3.26
x_8	9.25	5.09

จงแสดงการจัดกลุ่มแบบเคมีน เมื่อกำหนดจำนวนกลุ่มที่ต้องการเท่ากับ 2

7. จากชุดข้อมูลในข้อ 6 สามารถสร้างเมทริกซ์ระยะทางแบบยุคลิดของข้อมูล ได้ดังนี้

$$
\begin{array}{c}
\\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \end{array} \\
\begin{bmatrix}
0.00 & 1.64 & 1.68 & 1.33 & 3.02 & 8.29 & 6.38 & 5.94 \\
1.64 & 0.00 & 1.19 & 0.67 & 1.72 & 9.37 & 7.10 & 7.09 \\
1.68 & 1.19 & 0.00 & 0.6 & 1.47 & 9.94 & 7.90 & 7.60 \\
1.33 & 0.67 & 0.60 & 0.00 & 1.70 & 9.47 & 7.35 & 7.14 \\
3.02 & 1.72 & 1.47 & 1.70 & 0.00 & 11.08 & 8.81 & 8.78 \\
8.29 & 9.37 & 9.94 & 9.47 & 11.08 & 0.00 & 2.93 & 2.36 \\
6.38 & 7.10 & 7.90 & 7.35 & 8.81 & 2.93 & 0.00 & 2.12 \\
5.94 & 7.09 & 7.60 & 7.14 & 8.78 & 2.36 & 2.12 & 0.00
\end{bmatrix}
\end{array}
$$

จงแสดงการจัดกลุ่มข้อมูลในข้อ 6 โดยใช้วิธีการรวมกลุ่มแบบลำดับชั้น เมื่อกำนดให้จำนวนกลุ่มข้อมูลที่ต้องการมีค่าเท่ากับ 2 และใช้การเชื่อมต่อสมบูรณ์ในการคำนวณระยะห่างระหว่างกลุ่มข้อมูล

8. จากชุดข้อมูลในข้อ 6 จงแสดงการจัดกลุ่มโดยใช้วิธีการดีบีสแกน เมื่อกำหนดค่าพารามิเตอร์ $minPts = 2$ และ $\varepsilon = 3$

9. จากวิธีการวิเคราะห์จัดกลุ่ม 3 วิธี ได้แก่

1) วิธีการจัดกลุ่มแบบเคมีน 2) วิธีการรวมกลุ่มแบบลำดับชั้น 3) วิธีการดีบีสแกน

จงจับคู่วิธีการวิเคราะห์จัดกลุ่มที่เกี่ยวข้องกับคุณลักษณะ ต่อไปนี้

a) การรวมข้อมูลเป็นกลุ่มจากกลุ่มที่มีขนาดเล็กเป็นกลุ่มที่มีขนาดใหญ่ขึ้นจนกระทั่งได้จำนวนกลุ่มข้อมูลตามที่ต้องการ
b) การจัดข้อมูลที่อยู่ในบริเวณที่มีความหนาแน่นของข้อมูลบริเวณเดียวกันให้อยู่ในกลุ่มข้อมูลเดียวกัน
c) การแบ่งกลุ่มข้อมูลโดยมีเป้าหมายให้ผลรวมระยะทางระหว่างข้อมูลในกลุ่มข้อมูลเดียวกันมีค่าน้อยที่สุด

d) ต้องกำหนดจำนวนกลุ่มข้อมูลที่ต้องการ เป็นค่าพารามิเตอร์
e) ต้องกำหนดค่ารัศมีที่ระบุขอบเขตของจุดข้อมูลเพื่อนบ้าน โดยมีจุดข้อมูลเป็นศูนย์กลาง เพื่อคำนวณหาความหนาแน่นของข้อมูล ณ แต่ละจุดข้อมูล
f) สามารถตรวจหาข้อมูลผิดปรกติได้

10. จากชุดข้อมูลรายการซื้อสินค้าของร้านขายผลไม้แห่งหนึ่ง แสดงดังตารางต่อไปนี้

	Green grapes	Avocado	Tomato	Corn
t_1	0	1	1	0
t_2	1	1	0	0
t_3	0	1	0	1
t_4	1	0	1	0
t_5	0	1	1	1
t_6	1	1	0	1
t_7	0	1	1	1
t_8	0	0	1	0

จงตอบคำถามต่อไปนี้

a) จงแจกแจงไอเท็มเซตทั้งหมดที่เป็นไปได้ของชุดข้อมูล
b) จงหาค่าสนับสนุนของแต่ละไอเท็มเซตทั้งหมดที่เป็นไปได้จากชุดข้อมูล
c) จงแสดงการค้นหาไอเท็มเซตถี่ (Frequent Itemset) ทั้งหมดจากชุดข้อมูล โดยใช้ขั้นตอนวิธีเอไพรออริ (Apriori Algorithm)
d) จงหาค่าสนับสนุนของกฎความสัมพันธ์ {Tomato} $\Rightarrow$ {Avocado}
e) จงหาค่าความเชื่อมั่นของกฎความสัมพันธ์ {Tomato} $\Rightarrow$ {Avocado}
f) จงแสดงการค้นหากฎความสัมพันธ์ (Association Rule) ที่น่าเชื่อถือทั้งหมดจากชุดข้อมูล
g) จงหาค่าลิฟท์ของกฎความสัมพันธ์ {Tomato} $\Rightarrow$ {Avocado}
h) นักศึกษาควรจัดรายการส่งเสริมการขาย โดยลดราคาอะโวคาโด (Avocado) ให้กับลูกที่ซื้อมะเขือเทศ (Tomato) ร่วมด้วย หรือไม่ จงอธิบายเหตุผลประกอบ