Linear Regression

สูตรสำคัญ:

- **ความชัน (m) :** $m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$
- **จุดตัดแกน Y (c) :** $c = \bar{y} - m\bar{x}$

โจทย์ข้อที่ 1.1

บริษัทขายไอศกรีมต้องการทำนายยอดขาย (ถ้วย) จากอุณหภูมิสูงสุดของวัน (องศาเซลเซียส) โดยมีข้อมูล 5 วันล่าสุดดังนี้

อุณหภูมิ (X)	**ยอดขาย (Y)**
25	150
30	200
32	230
28	180
35	250

คำสั่ง:

1. จงหาสมการ Linear Regression ***(y=mx+c)*** จากข้อมูลข้างต้น
2. ถ้าวันนี้อุณหภูมิ 33 องศาเซลเซียส คาดว่าจะขายไอศกรีมได้กี่ถ้วย?

เฉลยข้อที่ 1.1

ขั้นตอนที่ 1: คำนวณค่าผลรวมต่างๆ

อุณหภูมิ (X)	**ยอดขาย (Y)**	X^2	XY
25	150	625	3750
30	200	900	6000
32	230	1024	7360
28	180	784	5040
35	250	1225	8750
∑x = 150	∑y = 1010	$\sum X^2$ = 4558	∑XY = 30900
n = 5			

ขั้นตอนที่ 2: คำนวณหาความชัน (m)

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$m = \frac{5(30900) - (150)(1010)}{5(4558) - (150)^2}$$

$$m = \frac{154500 - 151500}{22790 - 22500}$$

$$m = \frac{3000}{290} \approx 10.34$$

ขั้นตอนที่ 3: คำนวณหาจุดตัดแกน Y (c)

ก่อนอื่นหาค่าเฉลี่ย:

$$\bar{x} = \frac{150}{5} = 30$$

$$\bar{y} = \frac{1010}{5} = 202$$

$$c = \bar{y} - m\bar{x}$$

$c = 202 - (10.34)(30)$

$c = 202 - 310.2 = -108.2$

ขั้นตอนที่ 4: สร้างสมการและทำนายผล

สมการคือ: $y = 10.34x - 108.2$

ทำนายยอดขายที่อุณหภูมิ 33 องศา:

$y = 10.34(33) - 108.2$

$y = 341.22 - 108.2 = 233.02$

คำตอบ: คาดว่าจะขายไอศกรีมได้ประมาณ **233 ถ้วย**

โจทย์ข้อที่ 1.2

ฟิตเนสแห่งหนึ่งต้องการวิเคราะห์ความสัมพันธ์ระหว่างจำนวนชั่วโมงที่ลูกค้าออกกำลังกายต่อสัปดาห์ (X) กับน้ำหนักที่ลดลงในหนึ่งเดือน (กก.) (Y)

ชั่วโมง/สัปดาห์ (X)	น้ำหนักที่ลด (Y)
3	1.5
5	2.0
2	1.0
6	3.0
4	2.2
7	3.5

คำสั่ง:

1. จงหาสมการ Linear Regression
2. หากลูกค้าออกกำลังกาย 8 ชั่วโมง/สัปดาห์ คาดว่าน้ำหนักจะลดลงกี่กิโลกรัม?

เฉลยข้อที่ 1.2

ขั้นตอนที่ 1: คำนวณค่าผลรวมต่างๆ

ชั่วโมง/สัปดาห์ (X)	**น้ำหนักที่ลด (Y)**	X^2	XY
3	1.5	9	4.5
5	2.0	25	10.0
2	1.0	4	2.0
6	3.0	36	18.0
4	2.2	16	8.8
7	3.5	49	24.5
∑x = 27	∑y = 13.2	$\sum X^2$ = 139	∑XY = 67.8
n = 6			

ขั้นตอนที่ 2: คำนวณหาความชัน (m)

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$m = \frac{6(67.8) - (27)(13.2)}{6(139) - (27)^2}$$

$$m = \frac{406.8 - 356.4}{834 - 729}$$

$$m = \frac{50.4}{105} = 0.48$$

ขั้นตอนที่ 3: คำนวณหาจุดตัดแกน Y (c)

ก่อนอื่นหาค่าเฉลี่ย:

$$\bar{x} = \frac{27}{6} = 4.5$$

$$\bar{y} = \frac{13.2}{6} = 2.2$$

$c = \bar{y} - m\bar{x}$

$c = 2.2 - (0.48)(4.5)$

$c = 2.2 - 2.16 = 0.04$

ขั้นตอนที่ 4: สร้างสมการและทำนายผล

สมการคือ: $y = 0.48x + 0.04$

ทำนายน้ำหนักที่ลดลงเมื่อออกกำลังกาย 8 ชั่วโมง/สัปดาห์:

$y = 0.48(8) + 0.04$

$y = 3.84 + 0.04 = 3.88$

คำตอบ: คาดว่าจะขายไอศกรีมได้ประมาณ **233 ถ้วย**

Decision Tree (Regression)

สูตรสำคัญ:

- **Standard Deviation (SD) :** $SD = \sqrt{\frac{\sum (y_i - \mu)^2}{n}}$
- **Standard Deviation Reduction (SDR) :** $SDR = SD_{parent} - (\omega_{left} SD_{left} + \omega_{right} SD_{right})$

โจทย์ข้อที่ 2.1

ต้องการสร้างโมเดลทำนาย "ราคามือสอง" (Y, หน่วยเป็นพันบาท) ของสมาร์ทโฟน โดยพิจารณาจาก "อายุการใช้งาน (เดือน)" (X1)

อายุ (X1)	**ราคา (Y)**
6	18
12	14
24	9
8	17
18	11

คำสั่ง: จงหาการแบ่งครั้งแรก (First Split) ที่ดีที่สุด โดยคำนวณค่า Standard Deviation Reduction (SDR) ของทุกจุดแบ่งที่เป็นไปได้

เฉลยละเอียดโจทย์ข้อ 2.1

เป้าหมาย: หาจุดแบ่งที่ดีที่สุดสำหรับข้อมูลทั้งหมด (Root Node)

ข้อมูลเริ่มต้น:

- **X1 (อายุ):** {6, 8, 12, 18, 24}
- **Y (ราคา):** {18, 17, 14, 11, 9} (เรียงตาม X1)
- **SD ของข้อมูลทั้งหมด (SD_{parent}):** ≈3.429
- **จำนวนข้อมูลทั้งหมด (N):** 5
- **จุดแบ่งที่เป็นไปได้:** 7, 10, 15, 21

การคำนวณสำหรับจุดแบ่งที่ 1: อายุ <= 7

- **กลุ่มซ้าย (Y):** {18}
  - N = 1, μ = 18
  - SD_{left}=0 (เพราะมีข้อมูลเดียว)
- **กลุ่มขวา (Y):** {17, 14, 11, 9}
  - N = 4, μ = (17+14+11+9)/4 = 12.75
  - $\sum(y_i-\mu)^2$
    =(17−12.75)2+(14−12.75)2+(11−12.75)2+(9−12.75)2=18.06+1.56+3.06+14.06=36.74
  - SD_{right} = $\sqrt{36.74/4}$ ≈3.03
- **SDR:** $3.429-[(\frac{1}{5}\times 0)+(\frac{4}{5}\times 3.03)]=3.429-2.424=1.005$

การคำนวณสำหรับจุดแบ่งที่ 2: อายุ <= 10

- **กลุ่มซ้าย (Y):** {18, 17}
  - N = 2, μ = 17.5
  - SD_{left} = $\sqrt{((18-17.5)^2+(17-17.5)^2)/2} = \sqrt{0.5/2} = 0.5$
- **กลุ่มขวา (Y):** {14, 11, 9}
  - N = 3, μ = (14+11+9)/3 = 11.33
  - SD_{right} = $\sqrt{((14-11.33)^2+(11-11.33)^2+(9-11.33)^2)/3} \approx 2.05$
- **SDR:** $3.429-[(\frac{2}{5}\times 0.5)+(\frac{3}{5}\times 2.05)] = 3.429-[0.2+1.23] = 1.999$

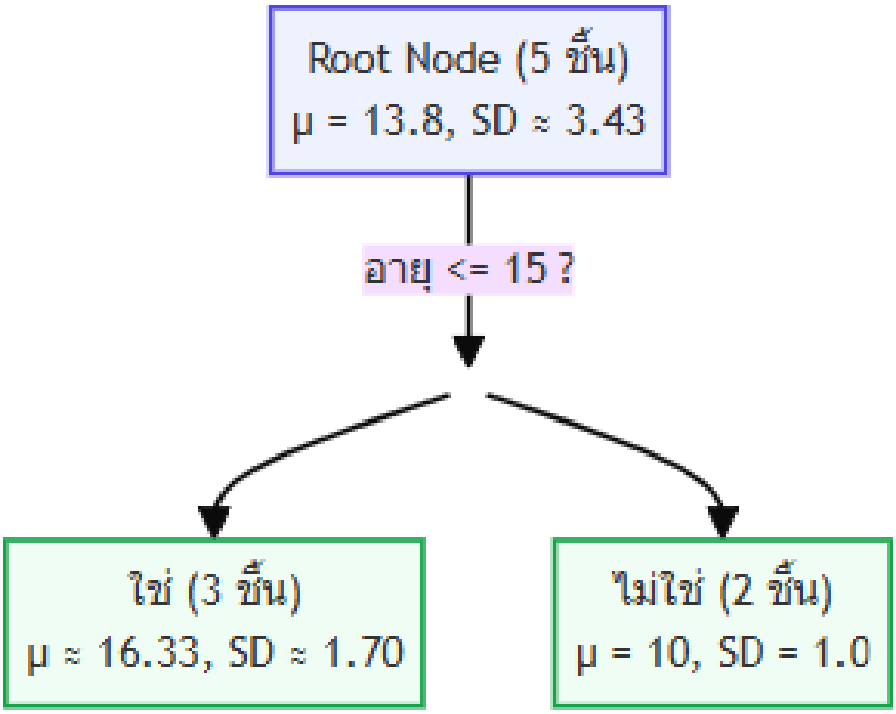
การคำนวณสำหรับจุดแบ่งที่ 3: อายุ <= 15 (จุดที่ดีที่สุด)

- **กลุ่มซ้าย (Y):** {18, 17, 14}
  - N = 3, μ = (18+17+14)/3 = 16.33
  - SD_{left} = $\sqrt{((18-16.33)^2+(17-16.33)^2+(14-16.33)^2/3} \approx 1.70$
- **กลุ่มขวา (Y):** {11, 9}
  - N = 2, μ = 10
  - SD_{right} = $\sqrt{((11-10)^2+(9-10)^2/2} = \sqrt{2/2} = 1$
- **SDR:** $3.429-[(\frac{3}{5}\times 1.70)+(\frac{2}{5}\times 1)] = 3.429-[1.02+0.4] = 2.009$

การคำนวณสำหรับจุดแบ่งที่ 4: อายุ <= 21

- **กลุ่มซ้าย (Y):** {18, 17, 14, 11}
  - N = 4, μ = 15
  - $SD_{left} = \sqrt{((18-15)^2+(17-15)^2+(14-15)^2+(11-15)^2/4} \approx 2.94$
- **กลุ่มขวา (Y):** {9}
  - N = 1, μ = 9
  - SD_{right} = **0**
- **SDR:** $3.429-[(\frac{4}{5}\times 2.94)+(\frac{1}{5}\times 0)] = 3.429-2.352 = 1.077$

สรุป: เมื่อเปรียบเทียบค่า SDR ทั้งหมด ค่าที่สูงสุดคือ **2.009** ซึ่งมาจากการแบ่งที่ อายุ <= 15



โจทย์ข้อที่ 2.2 (โจทย์ท้าทาย)

บริษัทเกมต้องการสร้างโมเดลทำนาย "คะแนนในเกม" (Y) ของผู้เล่น โดยอ้างอิงจาก "ชั่วโมงที่เล่น" (X1) และ "เลเวลผู้เล่น" (X2) **เงื่อนไข:** หยุดแบ่ง Node (สร้าง Leaf) ก็ต่อเมื่อ Node นั้นมีข้อมูลน้อยกว่าหรือเท่ากับ 3 ชิ้น

ชั่วโมงที่เล่น (X1)	เลเวลผู้เล่น (X2)	คะแนนในเกม (Y)
5	10	1200
15	25	3500
20	30	4500
2	5	500
8	15	1800
25	40	6000
12	20	2800
18	35	4000

คำสั่ง:

1. จงสร้าง Decision Tree จากข้อมูลทั้งหมดให้สมบูรณ์ตามขั้นตอน (แสดงการคำนวณเพื่อหาจุดแบ่งที่ดีที่สุดในแต่ละ Node)
2. วาดแผนผังต้นไม้ (Decision Tree) ที่สร้างเสร็จแล้ว
3. หากมีผู้เล่นใหม่ที่มี **ชั่วโมงที่เล่น 10 ชั่วโมง** และ **เลเวล 18** จงทำนายคะแนนของเขา

เฉลยละเอียดโจทย์ข้อ 2.2

เป้าหมาย: สร้าง Tree ทั้งหมดจนจบ โดยเริ่มจาก Root Node

รอบที่ 1: การแบ่งที่ Root Node

- **ข้อมูล:** 8 ชิ้น
- **SD_{parent}:** ≈1765.0
- **จุดแบ่งที่ดีที่สุด (คำนวณเหมือนข้อ 2.1 แต่มี 2 features):** คือ ชั่วโมงที่เล่น (X1) <= 13.5 เพราะให้ค่า **SDR สูงสุด ≈1147.2**
- **ผลลัพธ์:** ข้อมูลถูกแบ่งเป็น 2 Node
    - **Node ซ้าย:** (X1 <= 13.5) มี 4 ชิ้น (ต้องแบ่งต่อ)
    - **Node ขวา:** (X1 > 13.5) มี 4 ชิ้น (ต้องแบ่งต่อ)

รอบที่ 2: การแบ่งที่ Node ซ้าย (X1 <= 13.5)

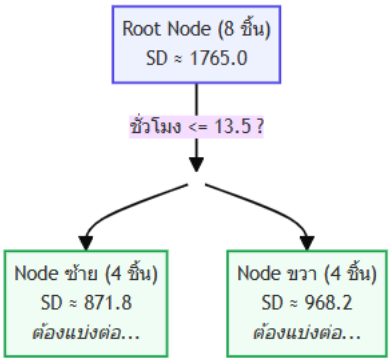
- **ข้อมูล:** { (5,10,1200), (2,5,500), (8,15,1800), (12,20,2800) }
- **SD_{parent} ของ Node นี้:** ≈871.8
- **จุดแบ่งที่เป็นไปได้:**
    - X1: 3.5, 6.5, 10
    - X2: 7.5, 12.5, 17.5
- **การคำนวณจุดแบ่งที่ดีที่สุดสำหรับ Node นี้:**
    - SDR ของ X2 <= 7.5: ≈390.9
    - SDR ของ X2 <= 12.5: ≈445.5 **(สูงสุด)**
    - SDR ของ X2 <= 17.5: ≈390.9

- **การตัดสินใจ:** เลือกแบ่งด้วย เลเวล (X2) <= 12.5
- **ผลลัพธ์:** Node ซ้ายถูกแบ่งเป็น 2 Leaf
    - **Leaf L-L (Y):** {1200, 500}. N=2 (<3). **หยุด**. ค่าทำนาย = (1200+500)/2 = **850**.
    - **Leaf L-R (Y):** {1800, 2800}. N=2 (<3). **หยุด**. ค่าทำนาย = (1800+2800)/2 = **2300**.

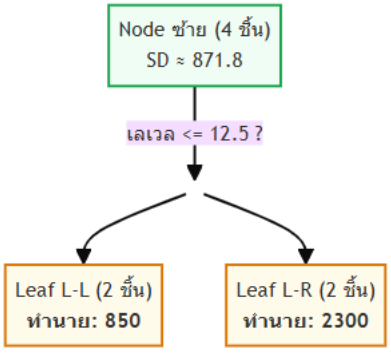
รอบที่ 3: การแบ่งที่ Node ขวา (X1 > 13.5)

- **ข้อมูล:** { (15,25,3500), (20,30,4500), (25,40,6000), (18,35,4000) }
- SD_{parent} **ของ Node นี้:** ≈968.2
- **จุดแบ่งที่เป็นไปได้:**
    - X1: 16.5, 19, 22.5
    - X2: 27.5, 32.5, 37.5
- **การคำนวณจุดแบ่งที่ดีที่สุดสำหรับ Node นี้:**
    - SDR ของ X2 <= 27.5: ≈434.6
    - SDR ของ X2 <= 32.5: ≈588.6 **(สูงสุด)**
    - SDR ของ X2 <= 37.5: ≈497.1
- **การตัดสินใจ:** เลือกแบ่งด้วย เลเวล (X2) <= 32.5
- **ผลลัพธ์:** Node ขวาถูกแบ่งเป็น 2 Leaf
    - **Leaf R-L (Y):** {3500, 4500}. N=2 (<3). **หยุด**. ค่าทำนาย = (3500+4500)/2 = **4000**.
    - **Leaf R-R (Y):** {6000, 4000}. N=2 (<3). **หยุด**. ค่าทำนาย = (6000+4000)/2 = **5000**.

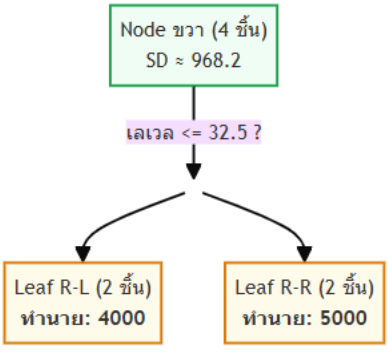
ขั้นตอนที่ 1: แบ่ง Root Node



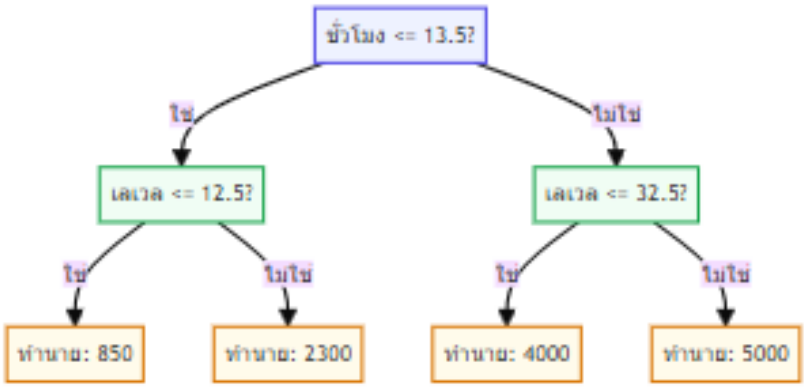
ขั้นตอนที่ 2: แบ่ง Node ซ้าย



ขั้นตอนที่ 3: แบ่ง Node ขวา



แผนผังต้นไม้ฉบับสมบูรณ์



K-Nearest Neighbors (K-NN)

สูตรสำคัญ:

- **ระยะห่างแบบยูคลิด (Euclidean Distance) :** $D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots}$

โจทย์ข้อที่ 3.1

นักวิเคราะห์สินเชื่อมีข้อมูลการอนุมัติสินเชื่อส่วนบุคคล โดยพิจารณาจาก "รายได้ต่อปี (แสนบาท)" (X1) และ "หนี้สินรวม (แสนบาท)" (X2)

ID	**รายได้ (X1)**	**หนี้สิน (X2)**	**ผลอนุมัติ (Y)**
P1	5	1	อนุมัติ
P2	6	3	อนุมัติ
P3	2	2	ไม่อนุมัติ
P4	3	4	ไม่อนุมัติ
P5	7	2	อนุมัติ
P6	4	5	ไม่อนุมัติ

คำสั่ง: ลูกค้าใหม่ (P_new) มี **รายได้ 6** แสนบาท และ **หนี้สิน 4** แสนบาท จงใช้ **K-NN (K=3)** ทำนายว่าลูกค้าคนนี้จะได้รับการอนุมัติหรือไม่?

เฉลยละเอียดโจทย์ข้อ 3.1 (K-NN, K=3)

เป้าหมาย: ทำนายว่าลูกค้าใหม่ P_new(6, 4) จะ "อนุมัติ" หรือ "ไม่อนุมัติ"

ข้อมูล:

- **อนุมัติ (A):** P1(5,1), P2(6,3), P5(7,2)
- **ไม่อนุมัติ (B):** P3(2,2), P4(3,4), P6(4,5)

ขั้นตอนที่ 1: คำนวณระยะห่างแบบยูคลิดจาก P_new(6, 4) ไปยังทุกจุด

สูตร: $D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots}$

ระยะทางถึง P1(5,1): $D = \sqrt{(6-5)^2 + (4-1)^2} = \sqrt{1^2 + 3^2} = \sqrt{1+9} = \sqrt{10} \approx 3.16$

ระยะทางถึง P2(6,3): $D = \sqrt{(6-6)^2 + (4-3)^2} = \sqrt{0^2 + 1^2} = \sqrt{0+1} = \sqrt{1} = 1$

ระยะทางถึง P3(2,2): $D = \sqrt{(6-2)^2 + (4-2)^2} = \sqrt{4^2 + 2^2} = \sqrt{16+4} = \sqrt{20} \approx 4.47$

ระยะทางถึง P4(3,4): $D = \sqrt{(6-3)^2 + (4-4)^2} = \sqrt{3^2 + 0^2} = \sqrt{9+0} = \sqrt{9} = 3.00$

ระยะทางถึง P5(7,2): $D = \sqrt{(6-7)^2 + (4-2)^2} = \sqrt{(-1)^2 + 2^2} = \sqrt{1+4} = \sqrt{5} \approx 2.24$

ระยะทางถึง P6(4,5): $D = \sqrt{(6-4)^2 + (4-5)^2} = \sqrt{2^2 + (-1)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.24$

ขั้นตอนที่ 2: จัดเรียงระยะทางจากน้อยไปมาก และเลือก 3 (K=3) อันดับแรก

อันดับ	ID	ผลอนุมัติ (Y)	ระยะทาง
1	P2	อนุมัติ	1.00
2	P5	อนุมัติ	2.24
3	P6	ไม่อนุมัติ	2.24
4	P4	ไม่อนุมัติ	3.00
5	P1	อนุมัติ	3.16
6	P3	ไม่อนุมัติ	4.47

ขั้นตอนที่ 3: ลงคะแนนเสียง (Majority Vote) จากเพื่อนบ้าน 3 อันดับแรก:

- **อนุมัติ:** 2 เสียง (จาก P2, P5)
- **ไม่อนุมัติ:** 1 เสียง (จาก P6)

สรุป: เสียงข้างมากคือ **"อนุมัติ"**

โจทย์ข้อที่ 3.2

มหาวิทยาลัยแห่งหนึ่งใช้ข้อมูล "เกรดเฉลี่ยตอน ม.ปลาย" (X1) และ "คะแนนสอบเข้า" (X2) เพื่อคัดกรองนักศึกษาที่มีแนวโน้มจะ "เรียนต่อจนจบ" หรือ "ลาออก"

ID	GPA (X1)	คะแนนสอบ (X2)	สถานะ (Y)
S1	3.8	85	เรียนจบ
S2	2.5	60	ลาออก
S3	3.5	90	เรียนจบ
S4	2.8	75	ลาออก
S5	3.2	80	เรียนจบ
S6	2.2	65	ลาออก
S7	3.9	95	เรียนจบ

คำสั่ง: นักเรียนใหม่ (S_new) มี **GPA 3.0** และ **คะแนนสอบ 70** จงใช้ **K-NN (K=5)** ทำนายสถานะของนักเรียนคนนี้

เฉลยละเอียดโจทย์ข้อ 3.2 (K-NN, K=5)

เป้าหมาย: ทำนายว่านักเรียนใหม่ S_new(3.0, 70) จะ "เรียนจบ" หรือ "ลาออก"

ขั้นตอนที่ 1: คำนวณระยะห่างจาก S_new(3.0, 70) ไปยังทุกจุด

ถึง S1(3.8, 85): $D = \sqrt{(3.0-3.8)^2 + (70-85)^2} = \sqrt{(-0.8)^2 + 5^2} = \sqrt{0.64 + 225} = \sqrt{225.64} \approx 15.02$

ถึง S2(2.5, 60): $D = \sqrt{(3.0-2.5)^2 + (70-60)^2} = \sqrt{0.5^2 + 10^2} = \sqrt{0.25 + 100} = \sqrt{100.25} \approx 10.01$

ถึง S3(3.5, 90): $D = \sqrt{(3.0-3.5)^2 + (70-90)^2} = \sqrt{0.5^2 + (-20)^2} = \sqrt{0.25 + 400} = \sqrt{400.25} \approx 20.01$

ถึง S4(2.8, 75): $D = \sqrt{(3.0-2.8)^2 + (70-75)^2} = \sqrt{0.2^2 + (-5)^2} = \sqrt{0.04 + 25} = \sqrt{25.04} \approx 5.00$

ถึง S5(3.2, 80):

$$D = \sqrt{(3.0-3.2)^2 + (70-80)^2} = \sqrt{(-0.2)^2 + (-10)^2} = \sqrt{0.04 + 100} = \sqrt{100.04} \approx 10.00$$

ถึง S6(2.2, 65): $D = \sqrt{(3.0-2.2)^2 + (70-65)^2} = \sqrt{0.8^2 + 5^2} = \sqrt{0.64 + 25} = \sqrt{25.64} \approx 5.06$

ถึง S7(3.9, 95):

$$D = \sqrt{(3.0-3.9)^2 + (70-95)^2} = \sqrt{(-0.9)^2 + (-25)^2} = \sqrt{0.81 + 625} = \sqrt{625.81} \approx 25.02$$

ขั้นตอนที่ 2: จัดเรียงระยะทางและเลือก 5 (K=5) อันดับแรก

อันดับ	ID	ผลอนุมัติ (Y)	ระยะทาง
1	S4	ลาออก	5.00
2	S6	ลาออก	5.06
3	S5	เรียนจบ	10.00
4	S2	ลาออก	10.01
5	S1	เรียนจบ	15.02
6	S3	เรียนจบ	20.01
7	S7	เรียนจบ	25.02

ขั้นตอนที่ 3: ลงคะแนนเสียง (Majority Vote)

- **ลาออก:** 3 เสียง (จาก S4, S6, S2)
- **เรียนจบ:** 2 เสียง (จาก S5, S1)

สรุป: เสียงข้างมากคือ **"ลาออก"**

4. Support Vector Machine (SVM)

โจทย์ข้อที่ 4.1

มีข้อมูล 2 คลาส คือ A (สีฟ้า) และ B (สีแดง)

- **คลาส A:** P1(2, 5), P2(3, 2)
- **คลาส B:** P3(6, 4), P4(7, 7)

มีคนเสนอเส้นแบ่ง (Hyperplane) H1 คือเส้นแนวดิ่ง x=4.5**ผิดพลาด! ไม่ได้ระบุชื่อไฟล์**

คำสั่ง:

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H1
2. เส้น H1 มี Support Vectors คือจุดใดบ้าง? และมี Margin กว้างเท่าใด?
3. จงหาเส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) และ Margin สูงสุดที่เป็นไปได้สำหรับข้อมูลชุดนี้

เฉลยละเอียดโจทย์ข้อ 4.1 (SVM)

เป้าหมาย: วิเคราะห์เส้นแบ่ง H1: $x - 4.5 = 0$

ข้อมูล:

- คลาส A: P1(2, 5), P2(3, 2)
- คลาส B: P3(6, 4), P4(7, 7)

ขั้นตอนที่ 1: คำนวณระยะห่างจากทุกจุดไปยังเส้นแบ่ง

สูตร: $d = \frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}} = \frac{|1x_0 + 0y_0 - 4.5|}{\sqrt{1^2 + 0^2}} = |x_0 - 4.5|$

จาก P1(2,5): $d = |2 - 4.5| = |-2.5| = 2.5$

จาก P2(3,2): $d = |3 - 4.5| = |-1.5| = 1.5$ (ใกล้สุดของคลาส A)

จาก P3(6,4): $d = |6 - 4.5| = |1.5| = 1.5$ (ใกล้สุดของคลาส B)

จาก P4(7,7): $d = |7 - 4.5| = |2.5| = 2.5$

ขั้นตอนที่ 2: หา Support Vectors และ Margin

- **Support Vectors** คือจุดที่อยู่ใกล้เส้นแบ่งที่สุดของแต่ละคลาส ซึ่งก็คือ **P2(3,2)** และ **P3(6,4)**
- **Margin** คือผลรวมของระยะทางจาก Support Vectors ไปยังเส้นแบ่ง: Margin = (ระยะทางจาก P2) + (ระยะทางจาก P3) = 1.5+1.5=3.0

ขั้นตอนที่ 3: วิเคราะห์ความเป็นเส้นแบ่งที่ดีที่สุด

- เส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) จะต้องอยู่กึ่งกลางระหว่าง Support Vectors พอดี
- จุดกึ่งกลางของพิกัด x ระหว่าง P2(3,2) กับ P3(6,4) คือ (3+6)/2=4.5
- เนื่องจากเส้นแบ่ง H1 (x=4.5) อยู่ ณ ตำแหน่งกึ่งกลางนี้พอดี **ดังนั้น H1 จึงเป็นเส้นแบ่งที่ดีที่สุด** และ Margin ที่คำนวณได้ (3.0) คือ Margin ที่กว้างที่สุดที่เป็นไปได้

โจทย์ข้อที่ 4.2

จากข้อมูลชุดเดิมในข้อ 4.1 มีคนเสนอเส้นแบ่งใหม่ H2 คือ x+y−8=0**ผิดพลาด! ไม่ได้ระบุชื่อไฟล์**

คำสั่ง:

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H2
2. เส้น H2 มี Support Vectors คือจุดใดบ้าง และ Margin กว้างเท่าใด?
3. เปรียบเทียบกับผลลัพธ์ในข้อ 4.1 เส้น H2 เป็นเส้นแบ่งที่ดีที่สุดหรือไม่ เพราะอะไร?

เฉลยละเอียดโจทย์ข้อ 4.2 (SVM)

เป้าหมาย: วิเคราะห์เส้นแบ่ง H2: $x + y - 8 = 0$

ขั้นตอนที่ 1: คำนวณระยะห่างจากทุกจุดไปยังเส้นแบ่ง

สูตร: $d = \frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}} = \frac{|1x_0 + 1y_0 - 8|}{\sqrt{1^2 + 1^2}} = \frac{|x_0 + y_0 - 8|}{\sqrt{2}}$

จาก P1(2,5): $d = \frac{|2 + 5 - 8|}{\sqrt{2}} = \frac{|-1|}{\sqrt{2}} \approx 0.707$ (ใกล้สุดของคลาส A)

จาก P2(3,2): $d = \frac{|3 + 2 - 8|}{\sqrt{2}} = \frac{|-3|}{\sqrt{2}} \approx 2.121$

จาก P3(6,4): $d = \frac{|6 + 4 - 8|}{\sqrt{2}} = \frac{2}{\sqrt{2}} \approx 1.414$ (ใกล้สุดของคลาส B)

จาก P4(7,7): $d = \frac{|7 + 7 - 8|}{\sqrt{2}} = \frac{|6|}{\sqrt{2}} \approx 4.243$

ขั้นตอนที่ 2: หา Support Vectors และ Margin

- Support Vectors ของเส้นนี้คือ P1(2,5) และ P3(6,4)
- Margin = (ระยะทางจาก P1) + (ระยะทางจาก P3) = 0.707+1.414=2.121

ขั้นตอนที่ 3: เปรียบเทียบและสรุป

- Margin ของ H2 (2.121) น้อยกว่า Margin ของ H1 (3.0) ที่เราหาได้ในข้อ 4.1
- นอกจากนี้ ระยะทางจากเส้น H2 ไปยัง Support Vector สองฝั่งก็ไม่เท่ากัน (0.707 vs 1.414) แสดงว่าเส้นยังไม่เป็นกลาง
- สรุป: เส้น H2 ไม่ใช่เส้นแบ่งที่ดีที่สุด เพราะยังไม่สามารถสร้างระยะห่างระหว่างกลุ่มได้กว้างที่สุดเท่าที่จะเป็นไปได้