

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

บริษัทขายไอศกรีมต้องการคำนวณยอดขาย (ถุง) จากอุณหภูมิสูงสุดของวัน (องศาเซลเซียส) โดยมีข้อมูล 5 วัน ล่าสุดดังนี้

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 คาดว่าจะขายไอศกรีมได้กี่ถุง?

$n=5$

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

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

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

$$m = \frac{3000}{290} \approx 10.34483$$

$$\begin{aligned} (\sum x) &= 25+30+32+28+35 = 150 \\ (\sum y) &= 150+200+230+180+250 = 1010 \\ (\sum xy) &= 3750+6000+7360+5040+8750 = 30900 \\ (\sum x^2) &= 625+900+1024+784+1225 = 4558 \end{aligned}$$

$$C = \bar{y} - m\bar{x} \quad \bar{y} = 202 \quad \bar{x} = 30$$

$$C = 202 - 10.34483 \cdot 30$$

$$C = 202 - 310.3449 = -108.3449$$

$$\therefore \text{สมการ คือ } Y = 10.34483X - 108.3449$$

$$\textcircled{2} \quad y = mx + c$$

$$y = 10.34 * 33 - 108.3$$

$$y = 341.22 - 108.3$$

$$y = 232.92 \approx 233$$

∴ \overline{m} 233 គឺជានៅក្នុងចំណែករវាង 232 និង 234

โจทย์ข้อที่ 1.2

พิฒนาสั่งหนึ่งต้องการวิเคราะห์ความสัมพันธ์ระหว่างจำนวนชั่วโมงที่ลูกค้าออกกำลังกายต่อสัปดาห์ (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 คาดว่าจะลดลงกี่กิโลกรัม? $n = 6$

①

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

$$m = \frac{6(67.8) - (27)(17.2)}{6(139) - (27)^2}$$

$$m = \frac{406.8 - 356.4}{934 - 729}$$

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

$$\begin{aligned} (\Sigma x) &= 3 + 5 + 2 + 6 + 4 + 7 = 27 \\ (\Sigma y) &= 1.5 + 2.0 + 1.0 + 3.0 + 2.2 + 3.5 = 13.2 \\ (\Sigma xy) &= 4.5 + 10 + 2 + 18 + 8.8 + 24.5 = 67.8 \\ (\Sigma x^2) &= 9 + 25 + 4 + 36 + 16 + 49 = 139 \end{aligned}$$

$$C = \bar{y} - m \bar{x} \quad \bar{y} = 2.2 \quad \bar{x} = 4.5$$

$$C = 2.2 - 0.48 \cdot 4.5$$

$$C = 2.2 - 2.16 = 0.04$$

$$\therefore \text{สมการลําบาก} \quad y = 0.48x + 0.04 \quad \times$$

②

$$y = mx + c$$

$$y = 0.48^* 8 + 0.04$$

$$y = 3.84 + 0.04$$

$$y = 3.88$$

∴ ຖ້າ ອົດຕະວັດທີ່ ດີວິForum ພົມ/ສັບຕະຍົກ ລາຄາທີ່ຈະນະຄຳຫຼັງຈິກ 3.88 ວົກ. *

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

$$\mu = 13.8 \quad SD_{root} \approx 3.42929$$

$$SD((y_1 - \mu)^2 \approx 17.64 / (y_2 - \mu)^2 \approx 0.04 / (y_3 - \mu)^2 \approx 23.04$$

$$(y_4 - \mu)^2 \approx 10.24 \quad (y_5 - \mu)^2 \approx 7.84$$

$$SDR(1/4) \text{ ถ้า } y_1 \leq 7 = 0.27969 \quad SDR(2/4) \text{ ถ้า } y_1 \leq 10 = 0.174594$$

$$SDR(3/4) \text{ ถ้า } y_1 \leq 15 = 0.17009 \quad SDR(4/4) \text{ ถ้า } y_1 \leq 21 = 0.27599 /$$

$$\therefore \text{ การแบ่งครึ่งแรก ที่ดีที่สุด คือ } y_1 \leq 21 = 0.27599 \quad \text{ } \#$$

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. วัดแผนผังต้นไม้ (Decision Tree) ที่สร้างเสร็จแล้ว
3. หากมีผู้เล่นใหม่ที่มี ชั่วโมงที่เล่น 10 ชั่วโมง และ เลเวล 18 จะทำนายคะแนนของเขา

①

$$\mu = 3,037.5 \quad SD_{root} \approx 1712.40876$$

$$SD \approx \sqrt{\frac{1}{14} \sum (y_i - \mu)^2} \approx \sqrt{\frac{1}{14} (3,376,406.25)} \approx 1712.40876$$

$$(y_1 - \mu)^2 \approx 1,661.38793 \quad (y_2 - \mu)^2 \approx 1,659.02467$$

$$(y_3 - \mu)^2 \approx 1,659.02467 \quad (y_4 - \mu)^2 \approx 1,658.66801$$

$$(y_5 - \mu)^2 \approx 1,658.66801 \quad (y_6 - \mu)^2 \approx 1,658.3258$$

$$(y_7 - \mu)^2 \approx 1,658.3258 \quad (y_8 - \mu)^2 \approx 1,657.66993$$

$$SDR(1/14) \text{ ชน. } \leq 3.5 = 1,661.38793$$

$$SDR(4/14) \text{ ชน. } \leq 13.5 = 1,659.02467$$

$$SDR(2/14) \text{ ชน. } \leq 6.5 = 1,659.02467$$

$$SDR(5/14) \text{ ชน. } \leq 16.5 = 1,658.66801$$

$$SDR(3/14) \text{ ชน. } \leq 10 = 1,656.02901$$

$$SDR(6/14) \text{ ชน. } \leq 19 = 1,658.3258$$

$$SDR(7/14) \text{ ชน. } \leq 22.5 = 1,657.66993$$

$$SDR(2/14) | level \leq 7.5 = 1667.38293, SDR(12/14) | level \leq 22.5 = 1658.66807$$

$$SDR(9/14) | level \leq 12.5 = 7659.09019 \quad SDR(13/14) | level \leq 32.5 = 1658.3258$$

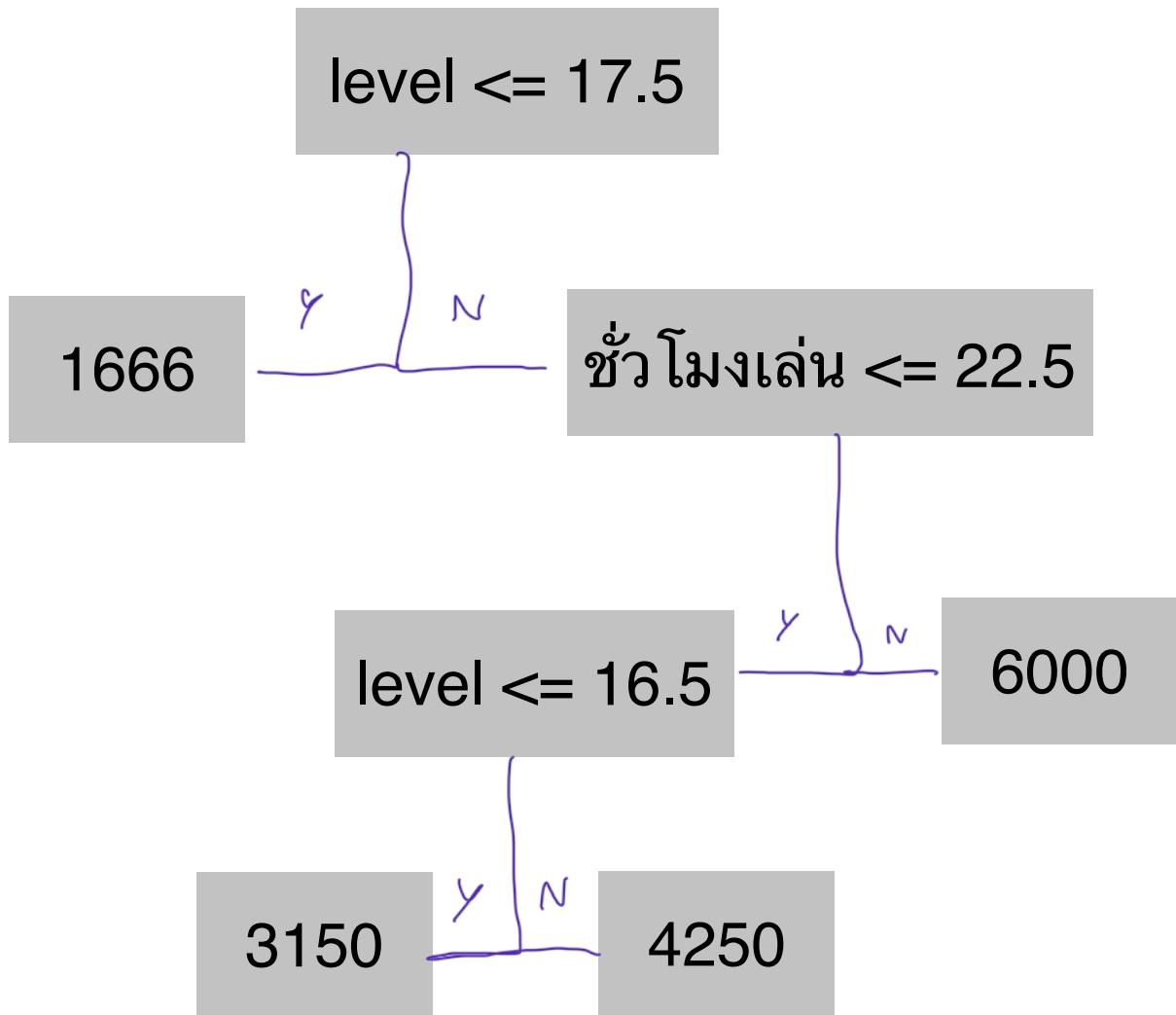
SDR (10/14) level <= 12.5 = 1659.02901

$SDR(13/14) |_{e \vee e} = 32F = 16FB 32FS$

$$SDR(17/14) \text{ level} \leq 32.5 = 1658.3258$$

SPR(74/74) level <= 37.5 = 1667.66993,

$$SDR(17/14) \text{ level } \zeta = 22.5 = 1659.02467$$



∴ ප්‍රේම් ඇත්තා නිසු. වේ 10 දා. මෙය බැංක් 18 නිවැන 11246 4250

K-Nearest Neighbors (K-NN)

สูตรสำคัญ:

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

โจทย์ข้อที่ 3.1

นักวิเคราะห์สินเชื่อมีข้อมูลการอนุมัติสินเชื่อส่วนบุคคล โดยพิจารณาจาก "รายได้ต่อปี (แสนบาท)" (X_1) และ "หนี้สินรวม (แสนบาท)" (X_2)

ID	รายได้ (X_1)	หนี้สิน (X_2)	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 จะใช้ K-NN ($K=3$) ทำนายว่าลูกค้าคนนี้จะได้รับการอนุมัติหรือไม่?

$$10 P_1 = 3.162 \quad 4$$

$$10 P_2 = 1 \quad 1 \quad 1$$

$$10 P_3 = 4.472 \quad 5$$

$$10 P_4 = 3 \quad 3$$

$$10 P_5 = 2.236 \quad 2 \quad 1$$

$$10 P_6 = 2.236 \quad 2 \quad 1$$

∴ ทำนายว่า ลูกค้าใหม่ จะรับการอนุมัติ \neq

โจทย์ข้อที่ 3.2

มหาวิทยาลัยแห่งหนึ่งใช้ข้อมูล "เกรดเฉลี่ยตอน ม.ปลาย" (X1) และ "คะแนนสอบเข้า" (X2) เพื่อคัดกรองนักศึกษา ที่มีแนวโน้มจะ "เรียนต่อจบ" 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 จะใช้ K-NN (K=5) คำนวณสถานะของนักเรียนคนนี้

$$S_1 = 15.021$$

$$S_2 = 10.012$$

$$S_3 = 20.006$$

$$S_4 = 5.004$$

$$S_5 = 10.002$$

$$S_6 = 5.064$$

$$S_7 = 25.016$$

∴ หัวข้อ รายงานผลการเรียน ระบุว่า จะลาออก 

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 คือเส้นแนวตั้ง $x=4.5$ ผิดพลาด! ไม่ได้ระบุชื่อไฟล์

คำสั่ง:

- จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H1
- เส้น H1 มี Support Vectors คือจุดใดบ้าง? และมี Margin กว้างเท่าใด?
- จงหาเส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) และ Margin สูงสุดที่เป็นไปได้สำหรับข้อมูลชุดนี้

①

$$P_1(2,5): r = \text{ระยะห่าง} = |2 - 4.5| = |-2.5| = 2.5$$

$$P_2(3,2): r = \text{ระยะห่าง} = |3 - 4.5| = |-1.5| = 1.5$$

$$P_3(6,4): r = \text{ระยะห่าง} = |6 - 4.5| = |1.5| = 1.5$$

$$P_4(7,7): r = \text{ระยะห่าง} = |7 - 4.5| = |2.5| = 2.5$$

②

$P_2(3,2)$ จากคลาส A ($r=1.5$)

$P_3(6,4)$ จากคลาส B ($r=1.5$)

$$\text{Margin กว้าง} = 1.5 + 1.5 = 3$$

③

$$x_2 = 3 \text{ และ } x_3 = 6$$

$$\text{ค่าเฉลี่ย } x = (3+6)/2 = 9/2 = 4.5$$

$$\therefore \text{Optimal Hyperplane } x < 4.5 \quad \text{Margin} = 1.5 + 1.5 = 3 \quad \times$$

โจทย์ข้อที่ 4.2

จากข้อมูลชุดเดิมในข้อ 4.1 มีคนเสนอเส้นแบ่งใหม่ H_2 คือ $x+y-8=0$ ผิดพลาด! ไม่ได้ระบุชื่อไฟล์

คำสั่ง:

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H_2
2. เส้น H_2 มี Support Vectors คือจุดใดบ้าง และ Margin กว้างเท่าใด?
3. เปรียบเทียบกับผลลัพธ์ในข้อ 4.1 เส้น H_2 เป็นเส้นแบ่งที่ดีที่สุดหรือไม่ เพราะอะไร?

① $P_1(2,5) : d_1 \approx 0.207$

$P_2(3,2) : d_2 \approx 2.121$

$P_3(6,4) : d_3 \approx 1.414$

$P_4(7,7) : d_4 \approx 4.243$

② $P_1(2,5) \approx 0.207$

$P_3(6,4) \approx 1.414$

Margin ≈ 2.121

③ \therefore เนื่องจากเส้น H_2 ที่ได้ที่สุดที่สุด หมายความว่า ระยะห่างของ Support Vectors ใกล้กัน