

## 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?

$$\sum x = 150 \quad \sum y = 1,010 \quad \sum x^2 = 4558 \quad \sum xy = 30,900$$

$$m = \frac{5(30,900) - (150)(1,010)}{5(4,558) - (22,500)}$$

$$m \approx 10.34483$$

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

$$C = 202 - 310.3449$$

$$C = -108.3449$$

$$y = mx + c$$

$$Y = (10.34483)(33) - 108.3449$$

$$Y \approx 233.03449 \text{ ถ้วย}$$

∴ ถ้าวันนี้อุณหภูมิ **33** องศาเซลเซียส คาดว่าจะขายได้ **233** ถ้วย

### โจทย์ข้อที่ 1.2

ฟิตเนสแห่งหนึ่งต้องการวิเคราะห์ความสัมพันธ์ระหว่างจำนวนชั่วโมงที่ลูกค้าออกกำลังกายต่อสัปดาห์ (X) กับ น้ำหนักที่ลดลงในหนึ่งเดือน (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?

$$\sum x = 27 \quad \sum y = 13.2 \quad \sum x^2 = 139 \quad \sum xy = 67.8$$

$$m = \frac{6(67.8) - (27)(13.2)}{6(139) - (729)}$$

$$m \approx 0.48$$

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

$$c = 2.2 - 2.16$$

$$c = 0.04$$

$$y = mx + c$$

$$y = (0.48)(8) + 0.04$$

$$y = 3.88$$

∴ หากลูกค้าออกกำลังกาย **8 ชั่วโมง/สัปดาห์** คาดว่าน้ำหนักจะลดลง **3.88** กิโลกรัม

## 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

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

$$\text{หา sd} \quad (y_1 - \mu)^2 \approx 17.64 \quad (y_2 - \mu)^2 \approx 0.04 \quad (y_3 - \mu)^2 \approx 23.04 \quad (y_4 - \mu)^2 \approx 10.24 \quad (y_5 - \mu)^2 \approx 7.84$$

$$SDR(1/4) \quad \text{อายุ} \leq 7 \quad = 0.27969$$

$$SDR(2/4) \quad \text{อายุ} \leq 10 \quad = 0.174594$$

$$SDR(3/4) \quad \text{อายุ} \leq 15 \quad = 0.170094$$

$$SDR(4/4) \quad \text{อายุ} \leq 21 \quad = 0.775994$$

∴ การแบ่งครั้งแรก (First Split) ที่ดีที่สุด คือ : อายุ ≤ 21 = 0.775994

โจทย์ข้อที่ 2.2 (โจทย์ท้าทาย)

บริษัทเกมต้องการสร้างโมเดลทำนาย "คะแนนในเกม" (Y) ของผู้เล่น โดยอ้างอิงจาก "ชั่วโมงที่เล่น" (X1) และ

"เลเวลผู้เล่น" (X2) **เงื่อนไข:** หยุตแบ่ง 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

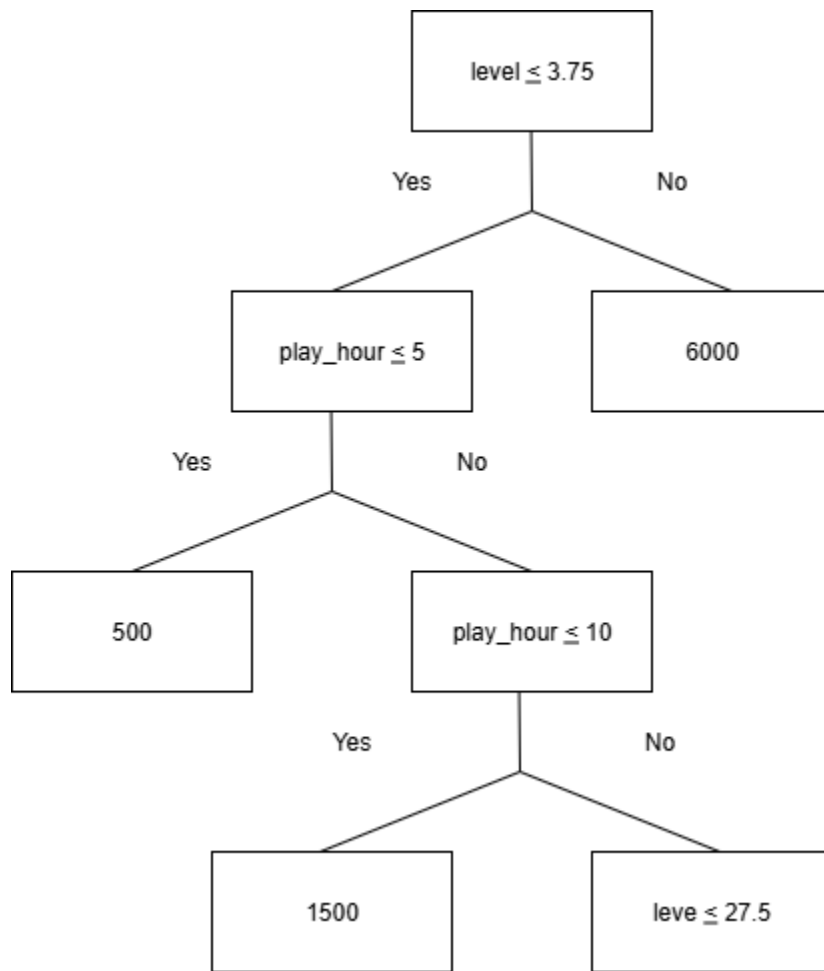
$$\mu = 3,037.5 \quad SD_{root} \approx 1712.40876$$

$$\text{หา sd} \quad (y_1 - \mu)^2 \approx 3,376,406.25 \quad (y_2 - \mu)^2 \approx 213,906.25 \quad (y_3 - \mu)^2 \approx 2,138,906.25$$

$$(y_4 - \mu)^2 \approx 6,438,906.25 \quad (y_5 - \mu)^2 \approx 1,531,406.25 \quad (y_6 - \mu)^2 \approx 8,776,406.25$$

$$(y_7 - \mu)^2 \approx 56,406.25 \quad (y_8 - \mu)^2 \approx 926,406.25$$

$$\begin{aligned} \text{SDR}(1/14) \text{ ชมเล่น} \leq 3.5 &= 1661.38793 & \text{SDR}(8/14) \text{ level} \leq 7.5 &= 1661.38793 \\ \text{SDR}(2/14) \text{ ชมเล่น} \leq 6.5 &= 1659.09019 & \text{SDR}(9/14) \text{ level} \leq 12.5 &= 1659.09019 \\ \text{SDR}(3/14) \text{ ชมเล่น} \leq 10 &= 1656.02901 & \text{SDR}(10/14) \text{ level} \leq 17.5 &= 1656.02901 \\ \text{SDR}(4/14) \text{ ชมเล่น} \leq 13.5 &= 1659.02461 & \text{SDR}(11/14) \text{ level} \leq 22.5 &= 1659.02461 \\ \text{SDR}(5/14) \text{ ชมเล่น} \leq 16.5 &= 1658.66801 & \text{SDR}(12/14) \text{ level} \leq 27.5 &= 1658.66801 \\ \text{SDR}(6/14) \text{ ชมเล่น} \leq 19 &= 1658.3258 & \text{SDR}(13/14) \text{ level} \leq 32.5 &= 1658.3258 \\ \text{SDR}(7/14) \text{ ชมเล่น} \leq 22.5 &= 1667.66993 & \text{SDR}(14/14) \text{ level} \leq 37.5 &= 1667.66993 \end{aligned}$$



ผู้เล่นใหม่ที่มีชั่วโมงที่เล่น 10 ชั่วโมง และ เลเวล 18 ทำนายคะแนนของเขาจะได้ 1500 คะแนน

## K-Nearest Neighbors (K-NN)

สูตรสำคัญ:

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

### โจทย์ข้อที่ 3.1

นักวิเคราะห์หุ้นเขามีข้อมูลการอนุมัติสินเชื่อส่วนบุคคล 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?

จุด P1 = 3.162

จุด P2 = 1

จุด P3 = 4.472

จุด P4 = 3

จุด P5 = 2.236

จุด P6 = 2.236

∴ ทำนายว่าลูกค้าคนนี้จะได้รับการอนุมัติ

### 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<sub>new</sub>) มี GPA 3.0 และ คะแนนสอบ 70 จงใช้ K-NN (K=5) ทำนายสถานะของนักเรียนคนนี้

S1 =15.021

S2 =10.012

S3 =20.006

S4 =5.004

S5 =10.002

S6 =5.064

S7 =25.016

∴ ทำนายสถานะของนักเรียนคนนี้ว่า 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 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

คำนวณหาระยะห่างจากทุกจุดไปยังเส้น H1

$$P1(2, 5): \text{ระยะห่าง} = |2 - 4.5| = |-2.5| = 2.5$$

$$P2(3, 2): \text{ระยะห่าง} = |3 - 4.5| = |-1.5| = 1.5$$

$$P3(6, 4): \text{ระยะห่าง} = |6 - 4.5| = |1.5| = 1.5$$

$$P4(7, 7): \text{ระยะห่าง} = |7 - 4.5| = |2.5| = 2.5$$

เส้น H2 มี Support Vectors คือจุดใดบ้าง และมี Margin กว้างเท่าใด ?

$$P2(3, 2) \text{ จากคลาส A (ระยะห่าง 1.5)}$$

$$P3(6, 4) \text{ จากคลาส B (ระยะห่าง 1.5)}$$

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

จงหาเส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) และ Margin สูงสุดที่เป็นไปได้สำหรับข้อมูลชุดนี้

$$x=3 \text{ และ } x=6$$

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

$$\text{ดังนั้น Optimal Hyperplane } x=4.5 \quad \text{Margin สูงสุด} = 1.5 + 1.5 = 3$$



## 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?

คำนวณหาระยะห่างจากทุกจุดไปยังเส้น  $H_2$

$P_1(2, 5)$ :  $d_1 \approx 0.707$

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

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

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

เส้น  $H_2$  มี Support Vectors คือจุดใดบ้าง Margin กว้างเท่าใด ?

$P_1(2, 5)$  ด้วยระยะห่าง  $\frac{1}{2}$

$P_3(6, 4)$  ด้วยระยะห่าง 2

Margin  $\approx 2.121$

เส้น  $H_2$  ไม่ใช่เส้นแบ่งที่ดีที่สุด เพราะ Margin ที่ได้จากเส้น  $H_2$  มีค่าน้อยกว่า Margin สูงสุดที่เป็นไปได้ซึ่งได้จากเส้น  $H_1$