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?

---

$\sum x = 150$ $\sum y = 1{,}010$ $\sum x^2 = 4{,}558$ $\sum xy = 30{,}900$

$$m = \frac{5(30{,}900) - (150)(1{,}010)}{5(4{,}558) - (22{,}500)} \qquad m \approx 10.34483$$

$c = \bar{y} - m\bar{x}$ C = 202-310.3449 C = -108.3449

$y = mx + c$ Y = (10.34483)(33)-108.3449 Y ≈ 233.03449 ถ้วย

∴ ถ้าวันนี้อุณหภูมิ 33 องศาเซลเซียส คาดว่าจะขายไอศกรีมได้ 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?

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

$m = \dfrac{6(67.8) - (27)(13.2)}{6(139) - (729)} \qquad m \approx 0.48$

$c = \bar{y} - m\bar{x} \qquad c = 2.2\text{-}2.16 \qquad c = 0.04$

$y = mx + c \qquad y = (0.48)(8)+0.04 \qquad y = 3.88$

∴ หากลูกค้าออกกำลังกาย 8 ชั่วโมง/สัปดาห์ คาดว่าน้ำหนักจะลดลง 3.88 KG

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 $SD_{root} \approx 3.42929$

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

SDR(1/4) อายุ<=7 = 0.27969
SDR(2/4) อายุ<=10 = 0.174594
SDR(3/4) อายุ<=15 = 0.170094
SDR(4/4) อายุ<=21 = 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:** 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

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

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

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

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

SDR(1/14) ชมเล่น<=3.5 = 1661.38793 SDR(8/14) level<=7.5 = 1661.38793

SDR(2/14) ชมเล่น<=6.5 = 1659.09019 SDR(9/14) level <=12.5 = 1659.09019

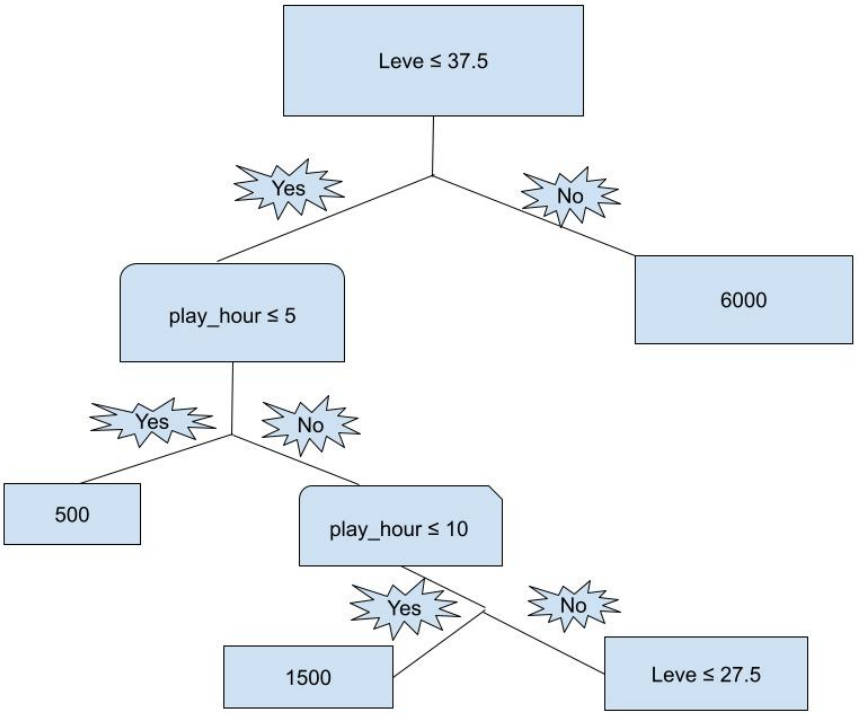
SDR(3/14) ชมเล่น<=10 = 1656.02901 SDR(10/14) level <=17.5 = 1656.02901

SDR(4/14) ชมเล่น<=13.5 = 1659.02461 SDR(11/14) level <=22.5 = 1659.02461

SDR(5/14) ชมเล่น<=16.5 = 1658.66801 SDR(12/14) level <=27.5 = 1658.66801

SDR(6/14) ชมเล่น<=19 = 1658.3258 SDR(13/14) level <=32.5 = 1658.3258

SDR(7/14) ชมเล่น<=22.5 = 1667.66993 SDR(14/14) level <=37.5 = 1667.66993




ผู้เล่นใหม่ที่มีชั่วโมงที่เล่น 10 ชั่วโมง และ เลเวล 18 ทำนายคะแนนของเขาจะได้ 1500 คะแนน

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?

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_new) 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

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H1

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

P2(3, 2): ระยะห่าง = $|3-4.5|=|-1.5|=1.5$

P3(6, 4): ระยะห่าง = $|6-4.5|=|1.5|=1.5$

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

2. เส้น H1 มี Support Vectors คือจุดใดบ้าง? และมี Margin กว้างเท่าใด?

P2(3, 2) จากคลาส A (ระยะห่าง 1.5)

P3(6, 4) จากคลาส B (ระยะห่าง 1.5)

Margin กว้าง = 1.5 + 1.5 = 3

3. จงหาเส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) และ Margin สูงสุดที่เป็นไปได้สำหรับข้อมูลชุดนี้

x=3 และ x=6

ค่าเฉลี่ยของ x = (3+6)/2=9/2=4.5

ดังนั้น Optimal Hyperplane x=4.5 Margin สูงสุด = 1.5 + 1.5 = 3

**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?

1. จงคำนวณหาระยะห่างจากทุกจุดไปยังเส้น H2

P1(2, 5): d1 ≈0.707
P2(3, 2): d2 ≈2.121
P3(6, 4): d3 ≈1.414
P4(7, 7): d4 ≈4.243

2. เส้น H2 มี Support Vectors คือจุดใดบ้าง และ Margin กว้างเท่าใด?

P1(2, 5) ด้วยระยะห่าง ½

P3(6, 4) ด้วยระยะห่าง 2

Margin ≈ 2.121

3. เปรียบเทียบกับผลลัพธ์ในข้อ 4.1 เส้น H2 เป็นเส้นแบ่งที่ดีที่สุดหรือไม่ เพราะอะไร?

เส้น H2 ไม่ใช่เส้นแบ่งที่ดีที่สุด เพราะ Margin ที่ได้จากเส้น H2 มีค่าน้อยกว่า Margin สูงสุดที่เป็นไปได้ซึ่งได้จากเส้น H1