


"

"

Data Mining

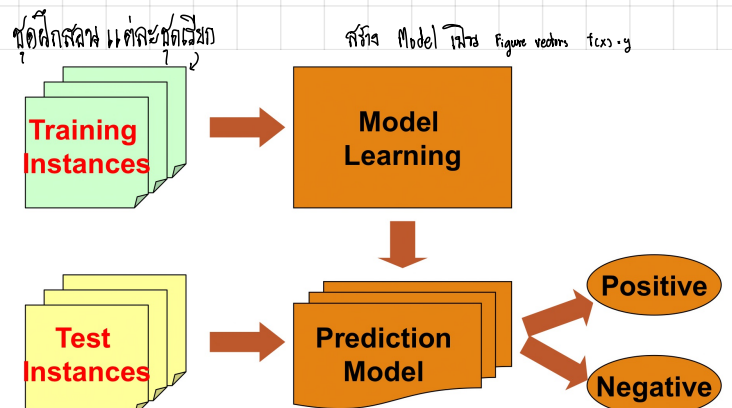


□ Supervised learning (classification)

คือกระบวนการที่จำแนกข้อมูลที่จะถูกส่งโดยหาวิธีที่เหมาะสมที่สุด

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

X Y



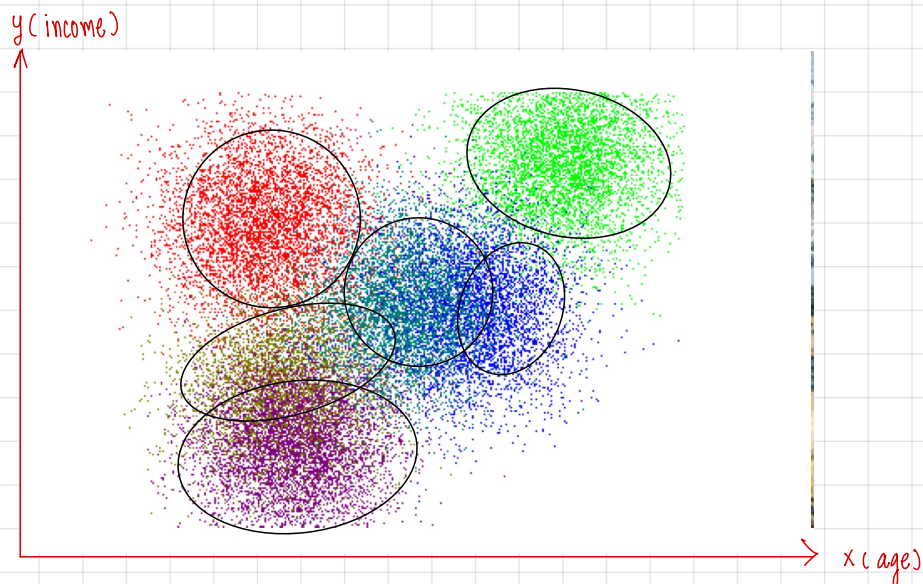
ทดสอบโมเดล (ประเมินการสลับ)

การหาค่าที่เหมาะสมที่สุดสำหรับแต่ละตัวอย่างที่ส่งเข้ามา
คือส่วน dataset เข้าไปแล้วให้ Model คำนวณหาว่าทำอย่างไร

□ Unsupervised learning (clustering)

เรียนรู้จาก Data โดยไม่มีผู้สอน

จัดกลุ่ม Data โดย Plot จุด บน x

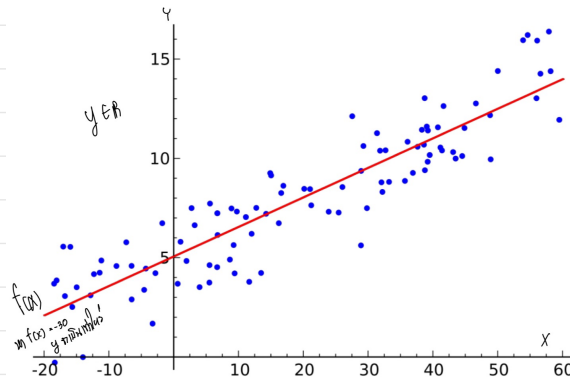


Classification

ค่า y เป็น class

→ เป็น category data เช่น ตั๋วเครื่องบิน ๑, ๒, ๓, ๔, ๕, ๖, ๗, ๘, ๙, ๑๐, ...

y จะถูกกำหนดเป็นตัวเลขเพื่อใช้ Regression



Model construction (เตรียมการสำหรับทำ)

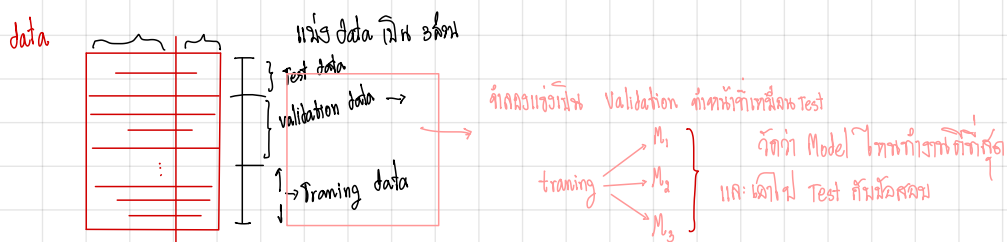
- เอา data มาสร้างเป็นโมเดล

Model Validation and Testing: (เตรียมการสำหรับ)

- ทดสอบโมเดล

Model Deployment:

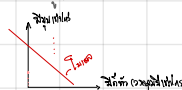
- เมื่อทดสอบผ่านแล้วจะเอาโมเดลไปใช้จริง



มอง x เป็นตัวช่วย

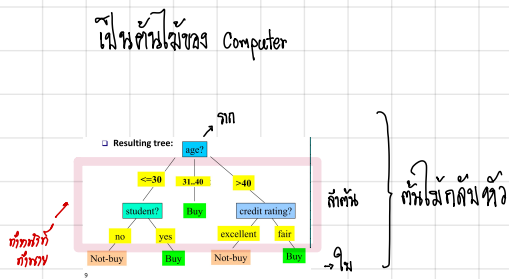
ก ก ก ก
ข ข ข ข

แต่ละตัวถูกแบ่งเป็น 4 เรขาคณิต



การแบ่งข้อมูลเป็น 4 กลุ่ม
ถ้ามี ๑๐๐ ข้อมูลจะแบ่งเป็น ๒๕
กลุ่ม

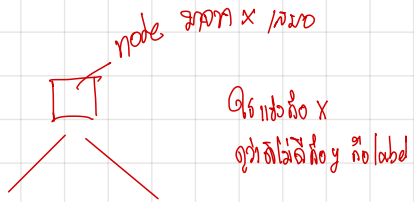
Decision tree construction:



สิ่งที่ได้ไม่เหมือนกัน อาจจะต่างกันที่ค่าที่ใส่

เมื่อใช้/ได้ค่าที่ใส่

node แทนค่าที่เราใส่ในโปรแกรม



สูตร

Expected information (entropy) needed to classify a tuple in D:

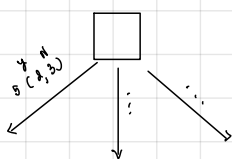
$$Info(D) = -\sum_{i=1}^m p_i \log_2(p_i)$$

Information needed (after using A to split D into v partitions) to classify D:

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$

Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$



top-down, recursive, divide-and-conquer manner

$$Info(D) = I(2, 3) = -\frac{2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right)$$

$$Info_{mean}(D) = \frac{2}{5} I(0, 2) + \frac{2}{5} I(1, 1) + \frac{1}{5} I(1, 0)$$

$$Info_{split}(D) = \frac{2}{5} I(2, 0) + \frac{3}{5} I(0, 3) \checkmark$$

$$Info_{max}(D) = \frac{2}{5} I(1, 2) + \frac{2}{5} I(1, 1)$$