Introduction to Supervised Learning

Petchara Pattarakijwanich

Introduction to Data Science, 2 September 2022

Goal of this week

- Intro to Supervised Learning
  - Types of Variables
  - Types of Methods
  - Test of Accuracy and Methodology
  - Goals of Supervised Learning

Example of Supervised Learning

Input

- GPA ม.ปลาย
- คะแนน TCAS

} Quantitative (อาจจะ correlate?)

- ประเมินจากลูกมารยาท (ดีมาก, ดี, ปกติ, ต้องปรับปรุง)

} Ordered Qualitative

- ช่วงเวลาเดือน ของแม่
- ตารีนของแม่
- โรงเรียนที่จบมา

} Qualitative

Output

- GPA ปีหน้า

Quantitative ⇒ Regression

- รอดปี 1 มั้ย?

Qualitative ⇒ Classification

⇒ สร้างโมเดลจาก input เพื่อทำนาย output

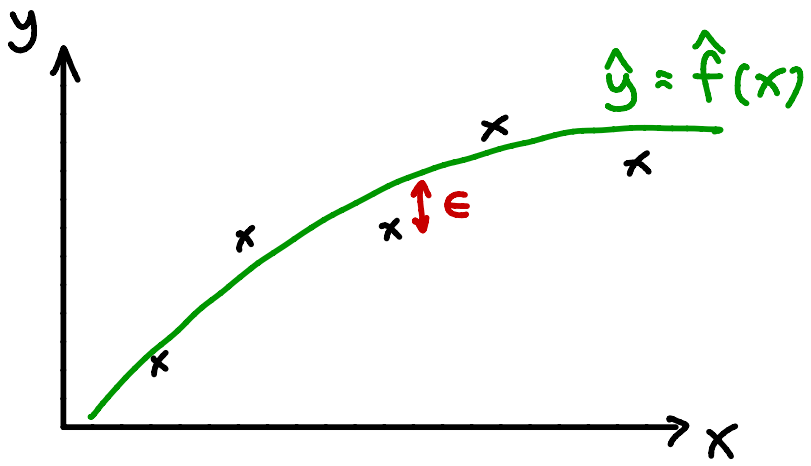
Mathematical Formulation

x = input , y = output

สมการรูปฟังก์ชัน $y = f(x) + \epsilon$

ϵ : Noise

f มีอยู่จริง แต่เราไม่รู้ ต้องประมาณ



ประมาณการรูปฟังก์ชัน $\hat{y} = \hat{f}(x)$

$\hat{y}$ คือค่าทำนาย (predicted value) ของ y ที่ตำแหน่ง x

$\hat{f}$ เป็น estimator ของ f

Irreduceable Noise ϵ

- ทำให้เป็น 0 ไม่ได้
- เกิดจาก - ความคลาดเคลื่อน
  - มี input อื่นที่ไม่ได้เก็บมา
  - x, y ในครั้ง รูปแบบ / เวลา / ครั้งทดลอง

$E(\epsilon) = 0$

$Var(\epsilon) = \sigma^2$

Types of Variables

By function

- x = input/independent/predictor/feature

- y = output/dependent/response/target

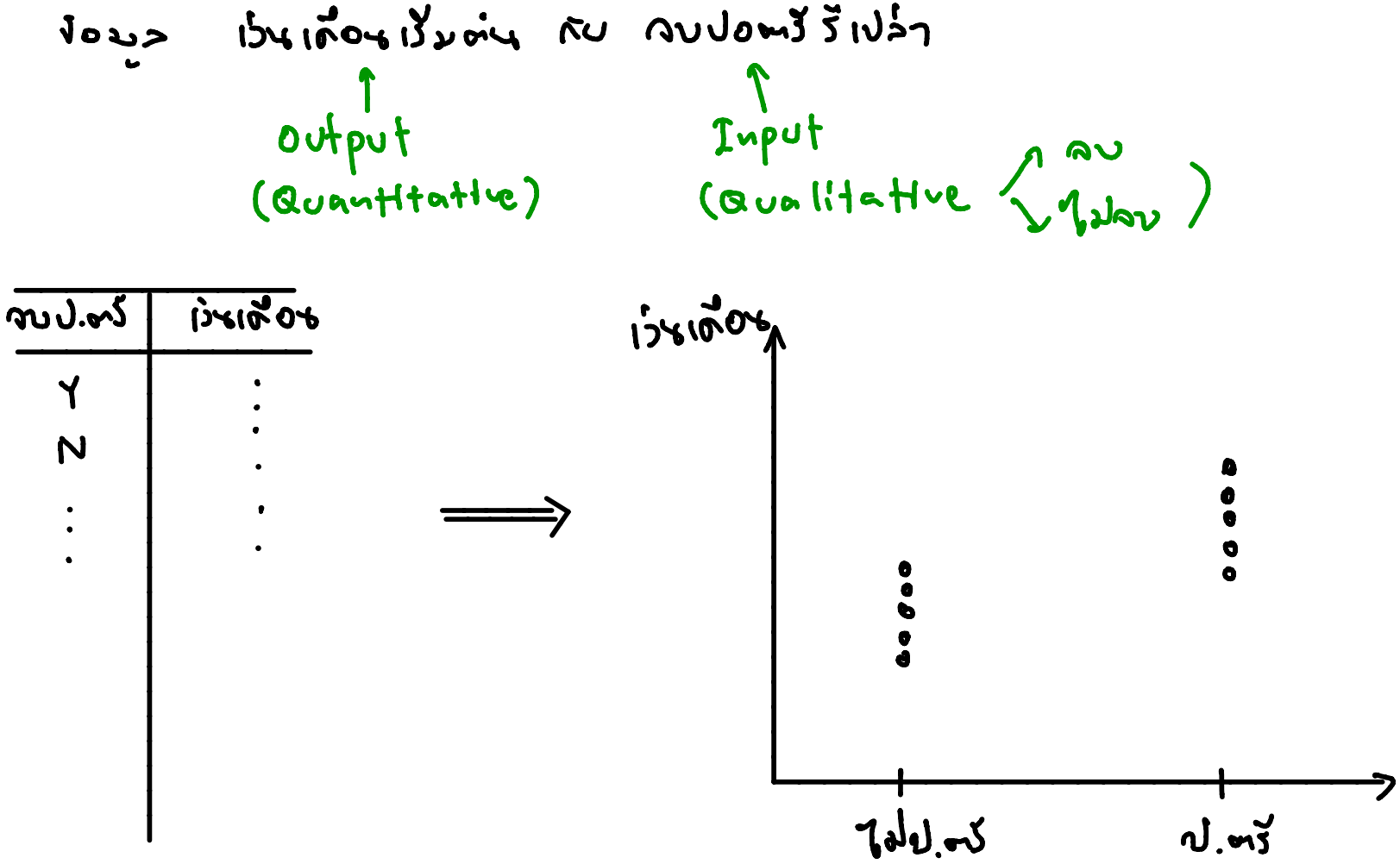
Types of Variables

By nature

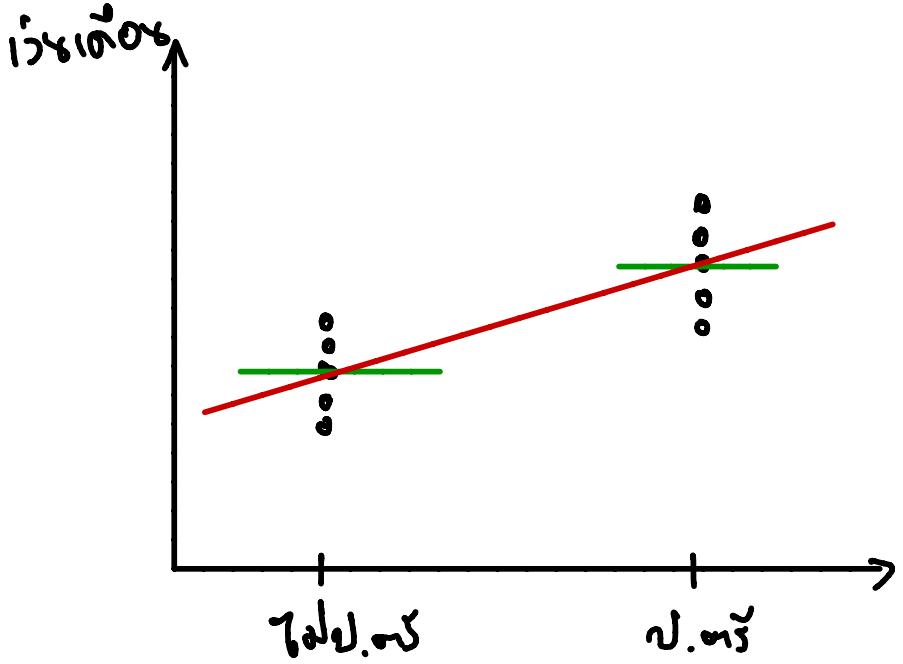
- Quantitative เป็นตัวเลข (จำนวนจริง/เต็ม)
  - เปรียบเทียบกันได้ ($>, <$ มีความหมายตัวเลข)
  - ต่างกัน $\Leftrightarrow$ ระยะห่างต่างกัน
- Qualitative (Categorical) เป็นกลุ่ม
  - เปรียบเทียบกันไม่ได้
  - 2 กลุ่ม (Binary) หรือ >2 กลุ่ม (multi-class)
- Ordered Quantitative

  Ex. ความปวด / พอใจ / ดี / ดีมาก

How to Deal with Qualitative Parameters?



How to Deal with Qualitative Parameters?



$(x = 0)$ $(x = 1)$

โมเดลง่ายสุด

$$\hat{f}(x) = \begin{cases} y_1, & \text{จบป.ตรี} \\ y_2, & \text{ไม่จบป.ตรี} \end{cases}$$

ไม่แยกกรณี

$\Rightarrow$ คำนวณ $\bar{x}$ ราคาเฉลี่ยแต่ละกลุ่ม

$\Rightarrow$ Regression

$$y = mx + c$$

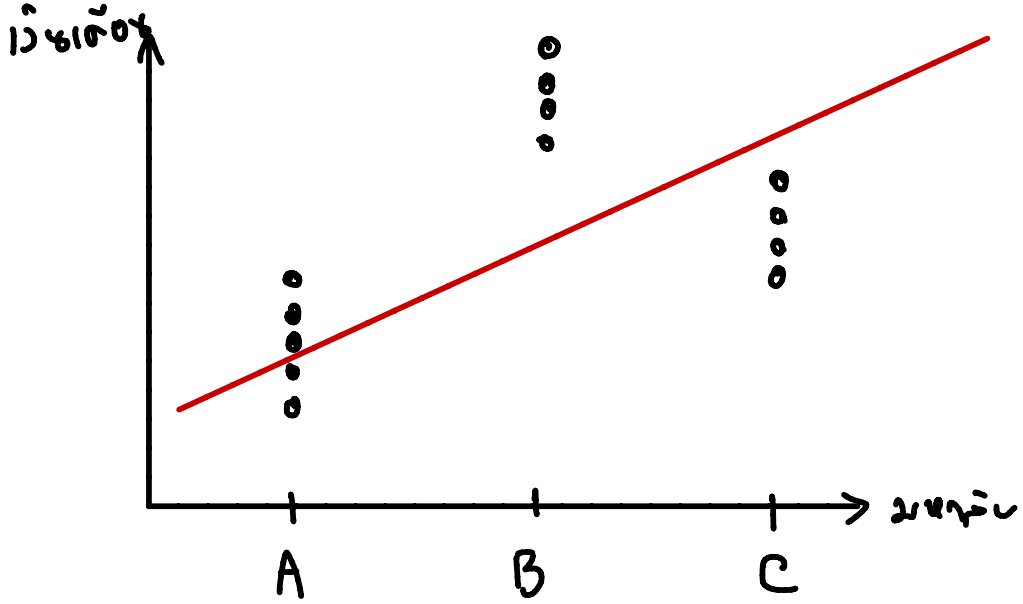
ตีความ?

- ไม่จบป.ตรี $y = c$
- จบป.ตรี $y = m + c$
- $y_{\text{ป.ตรี}} - y_{\text{ไม่ป.ตรี}} = m$

How to Deal with Qualitative Parameters?

ข้อมูล เงินเดือน vs มหาลัยที่จบมา (A,B,C)

มหาลัย	เงินเดือน
A	⋮
B	⋮
C	⋮
⋮	



(x=0) (x=1) (x=2)

วิธีนี้ใช้ไม่ได้!

How to Deal with Qualitative Parameters?

ข้อมูล เงินเดือน vs มหาลัยที่จบมา (A,B,C)

นิยาม $x_1 = \begin{cases} 0 & , \text{ A} \\ 1 & , \text{ ไม่ A} \end{cases}$

$x_2 = \begin{cases} 0 & , \text{ B} \\ 1 & , \text{ ไม่ B} \end{cases}$

มหาลัย	x_1	x_2
A	0	1
B	1	0
C	1	1

พิจารณาโมเดล

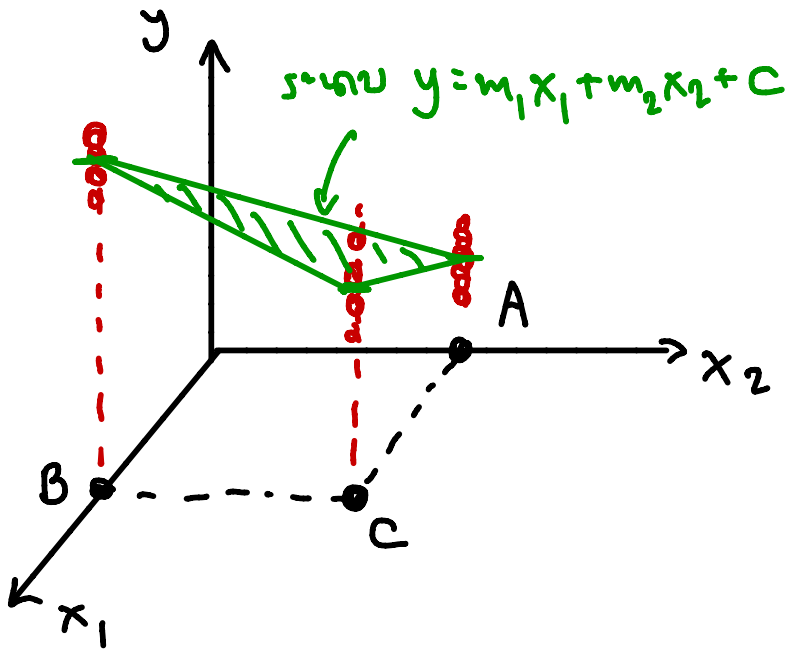
$$y = m_1 x_1 + m_2 x_2 + c$$

ดังนี้

มหาลัย A : $y = m_2 + c$

มหาลัย B : $y = m_1 + c$

มหาลัย C : $y = m_1 + m_2 + c$



Types of Methods

By output

- Regression Output = Quantitative

  Ex. พลัง 1

- Classification Output = Qualitative

  Ex. ดี/ไม่ดี

Types of Methods

By technique $y = f(x) + \epsilon$

- Parametric
  - สมมติรูปร่างของ f เช่น
  - เทคนิค function fitting
  - ข้อดี : ใช้ข้อมูลน้อยๆ
  - ข้อเสีย : ขึ้นกับว่าเดา f แม่นแค่ไหน
- Non-parametric
  - ยังไม่กำหนดรูปร่าง f
  - f ถูกขับเคลื่อนด้วยข้อมูล
  - ข้อดี : ไม่ขึ้นกับรมมติฐาน
  - ข้อเสีย : ใช้ข้อมูลเยอะ

How to Measure Accuracy?

- Regression

MSE = Mean Squared Error

$= \frac{1}{n} \sum (y - \hat{y})^2$ ⇒ ยิ่ง = ผิดมาก

"Loss Function"

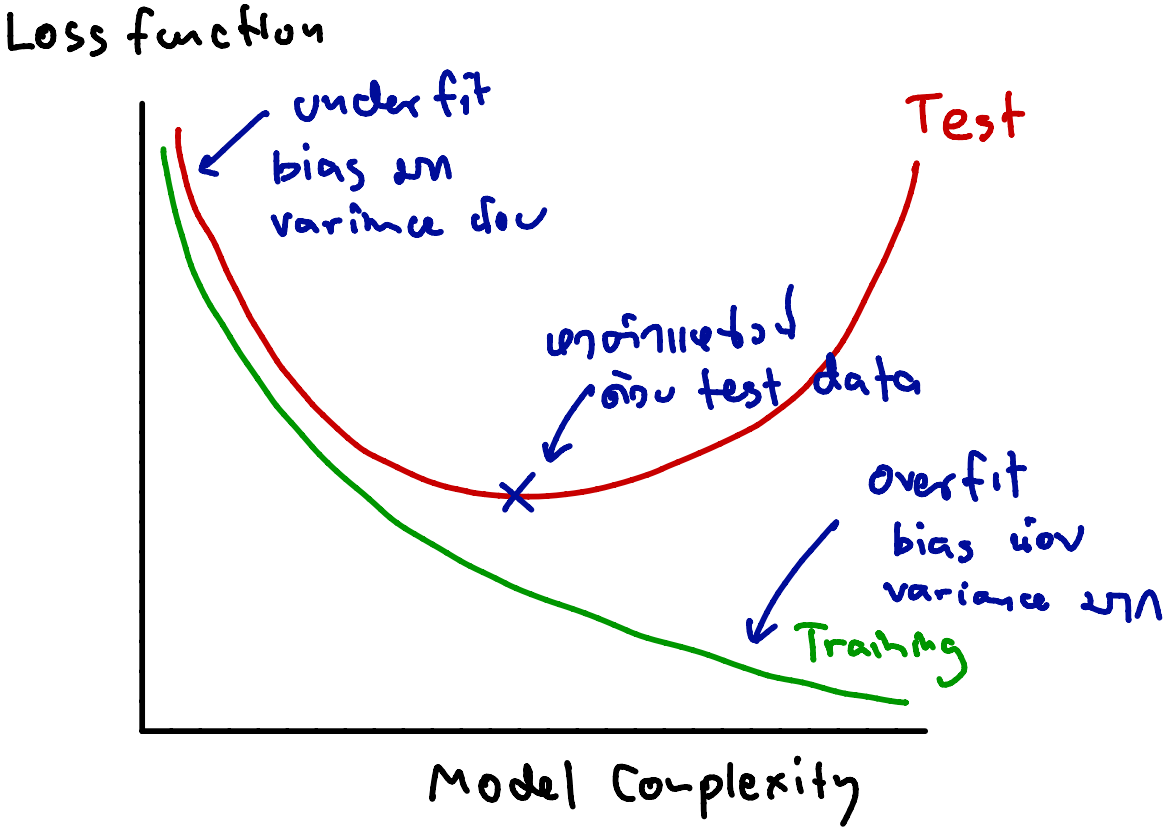
ต้องทำกับ test data

- Classification

Error Rate = $\frac{\text{จำนวนครั้งที่ทายผิด}}{\text{จำนวนครั้งทั้งหมด}}$ ยิ่ง = ทายแย่

Confusion Matrix (พจนานุกรม)

Bias-Variance Tradeoff



Methodology

- Training
  รู้ input/output ต้องการหา $\hat{f}$
- Testing (validation)
  รู้ input/output ต้องการรู้ผล $\hat{f}$ [$\hat{y} = \hat{f}(x)$ ⇒ เทียบ $y, \hat{y}$]
- Application
  รู้ input ไม่รู้ output ต้องการ $\hat{f}$ ทำนาย $\hat{y}$

Goals of Supervised Learning

- Prediction
  - ทำนาย y จาก x
  - วัดผลจากความแม่นยำ (Accuracy ฯลฯ)
  - ถือ $\hat{f}$ เป็น Black box (ไม่จำเป็นต้องรู้ว่าข้างในทำงานยังไง)
- Inference
  - เพื่อ <u>เข้าใจ</u> ความสัมพันธ์ระหว่าง x, y
    (Ex. ตัวแปรไหนเกี่ยวข้องกับ y บ้าง? ตัวแปรไหนสำคัญที่สุด?)
  - ต้องการโมเดลที่เข้าใจได้ อธิบายได้