

Resampling Method

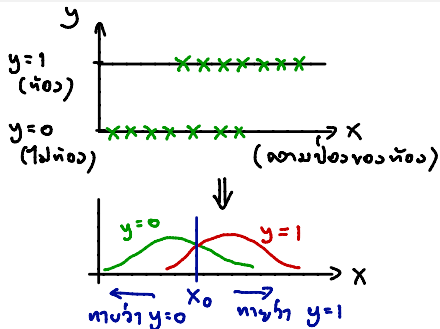
Petchara Pattarakijwanich

Introduction to Data Science, 30 September 2022

Goal of this week

- (Confusion matrix & ROC curve)
- Cross Validation Methods
 - Validation Set (Train & Test)
 - Leave-One-Out Cross Validation (LOOCV)
 - K-Fold Cross Validation
- Bootstrap Resampling

Sensitivity-Specificity Tradeoff

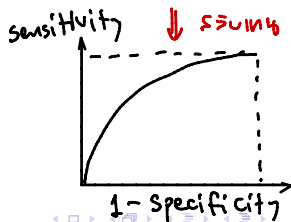
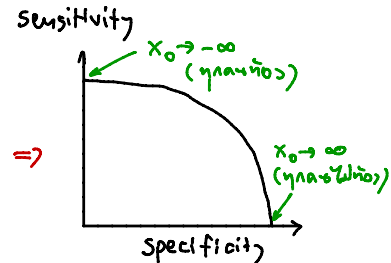
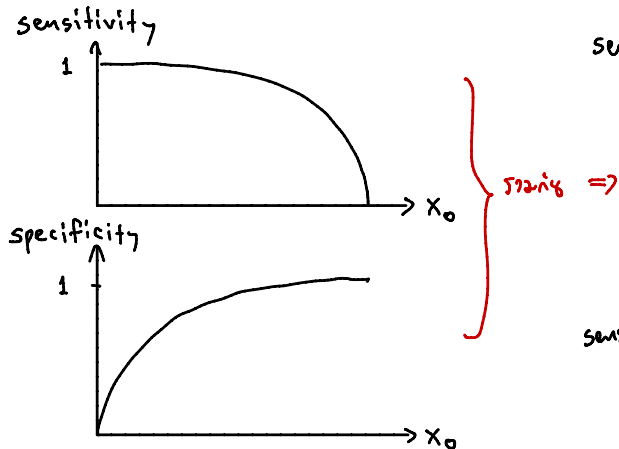


		predicted	
		+	-
Actual	+	OK	ความน่าเชื่อถือ
	-	ไม่จริง	OK

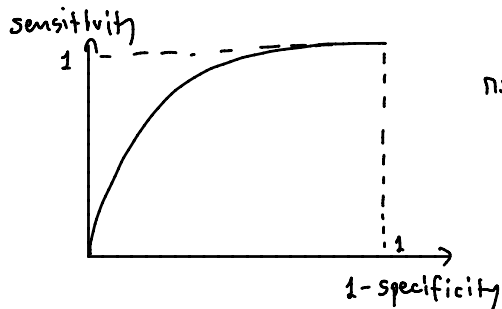
ความผิดพลาด FN รัฐบาล
 \Rightarrow รัฐบาลไม่จริง-ทาส่งไม่จริง
 \Rightarrow ความน่าเชื่อถือ x_0
 \Rightarrow sensitivity สูง
 specificity ต่ำ

ความผิดพลาด FP รัฐบาล
 \Rightarrow รัฐบาลไม่จริง-ทาส่งจริง
 \Rightarrow ความน่าเชื่อถือ x_0
 \Rightarrow sensitivity ต่ำ
 specificity สูง

၂၀။ x_0 ချိတ် $NS_0 \Rightarrow$ အကွေး confusion matrix (ဖော်ပြချက် x_0)
 \Rightarrow အကွေး sensitivity & specificity ဖော်ပြချက် x_0



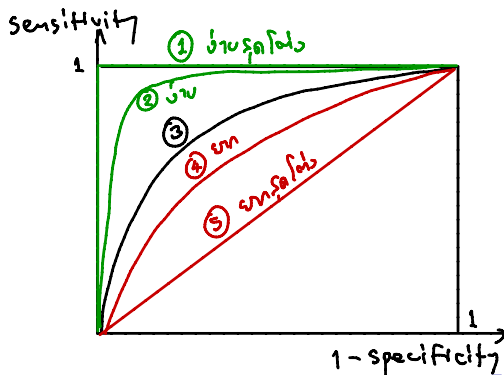
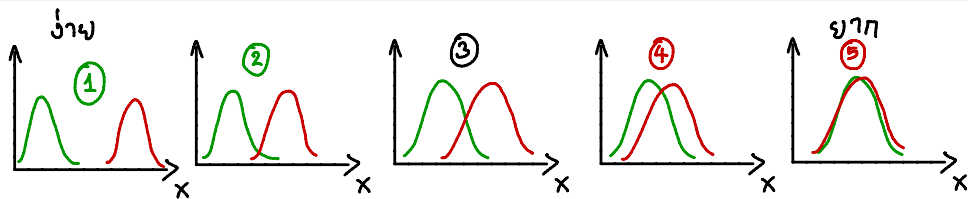
Receiver Operating Characteristic (ROC) Curve



ROC curve מודל מיושם

הקצאת כסף לטובת המחקר ולמחקר

Receiver Operating Characteristic (ROC) Curve



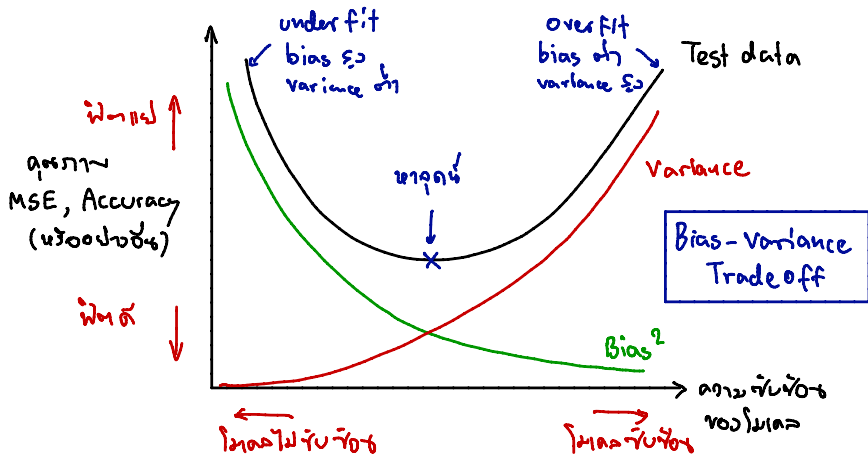
พื้นที่ใต้ ROC curve
ยิ่งใกล้ 1 ยิ่งดี

Validation Set

วิธีหาค่าที่ดีที่สุด

Training 50% + Test 50%

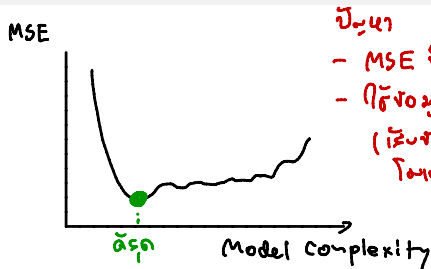
Fit model \Rightarrow วัดการ generalize



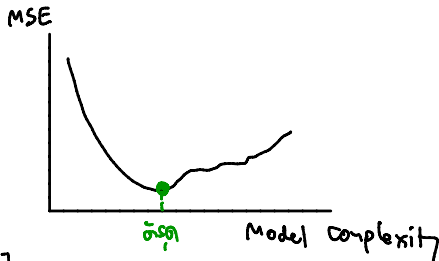
Validation Set



⇒



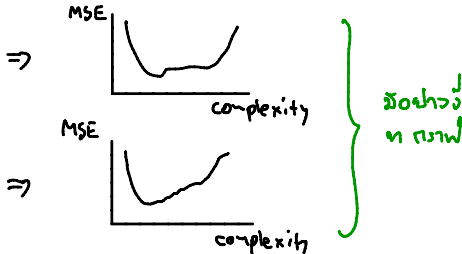
⇒



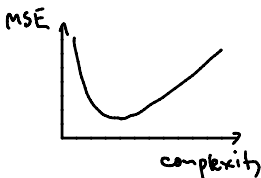
[ใช้ค่า training & test (11.5%)]

Leave-One-Out Cross Validation (LOOCV)

ចែក n ទិន្នន័យ \Rightarrow ប្រើ $n-1$ ទិន្នន័យ + test 1 ទិន្នន័យ



- កែប្រែទិន្នន័យ $n-1$ ទិន្នន័យ \Rightarrow ធ្វើវិញ
- ធ្វើបែបនេះច្រើនដង
- យកមធ្យម



លទ្ធផល

- ការប្រើ MSE តាមការប្រើប្រាស់
- ប្រើ ទិន្នន័យ $n-1$ ទិន្នន័យ (training $n-1$ ទិន្នន័យ)

លទ្ធផល

- ប្រើ (ទិន្នន័យ n ទិន្នន័យ complexity n ទិន្នន័យ)
- \Rightarrow ប្រើ ការប្រើប្រាស់

K-Fold Cross Validation

ឯង $k = 5, 10$

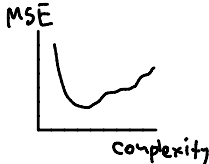
→ ឯង ចែកទិន្នន័យ k ដង

ក្នុង $k-1$ ដង ធ្វើ training + 1 ដង ធ្វើ test
ដូច្នេះបាន k លទ្ធផល

$k = 5$

Test (20%)
Train (80%)

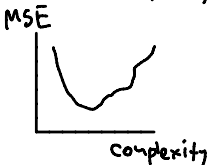
⇒



ឯង k ដង

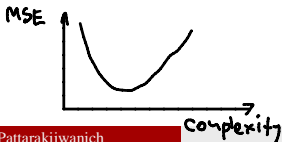
Test
Train

⇒



៖
៖
៖

ក្នុង k ដង \Rightarrow បាន



បំណង បំណង

(ការពិនិត្យ validation ឬ LOOCV)

- វាស់ស្ទង់ លំហូរ ឯង
 - វាស់ស្ទង់ លំហូរ ឯង
 - លំហូរ ឯង
- (ឯង ឯង $k \times m$ ដង)

K-Fold Cross Validation

Leave-one-out CV = n-fold CV

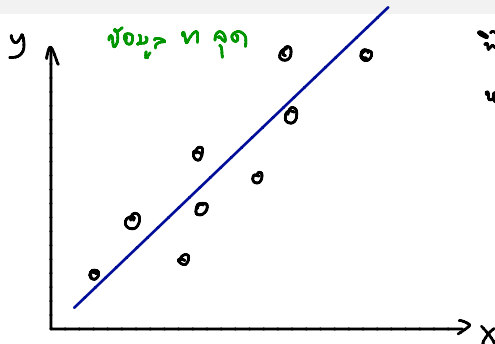
Validation set = $\frac{1}{2}$ [2-fold CV]

↑
1/2 for train & test

Summary of Cross Validation Methods

	Validation set 50% + 50%	LOOCV	K-fold CV
<p>วัดผล model 1 ครั้ง?</p> <p>(ง่าย → Objective)</p>	ง่าย	ง่ายและ	ง่ายพอ
<p>ใช้ข้อมูลทั้งหมด</p> <p>(ใช้ข้อมูลเพื่อ model 1 ครั้ง ⇒ model 1)</p>	<p>50%</p> <p>แบ่งข้อมูล</p>	<p>$n-1$ ครั้ง</p> <p>ครั้ง</p>	<p>$\frac{k-1}{k}$ รอบ</p> <p>ครั้ง</p>
ความเร็ว	เร็ว $(1 \times n_{\text{model}})$	ช้า $(n_{\text{data}} \times n_{\text{model}})$	กลาง $(k \times n_{\text{model}})$

Bootstrap Resampling



$$\hat{y} = mx + c$$

หาพารามิเตอร์ m, c ?

- ทฤษฎี
- **Bootstrap**

↓
ข้อมูล n จุด
แบบสุ่มซ้ำ



พารามิเตอร์ $\rightarrow m_1, c_1$

↓
ข้อมูล n จุด
แบบสุ่มซ้ำ



พารามิเตอร์ m_2, c_2

↓
ข้อมูล N จุด ($N \rightarrow \infty$)
 $m_1, m_2, \dots, m_N \Rightarrow \bar{m}, \sigma_m$
 $c_1, c_2, \dots, c_N \Rightarrow \bar{c}, \sigma_c$

