



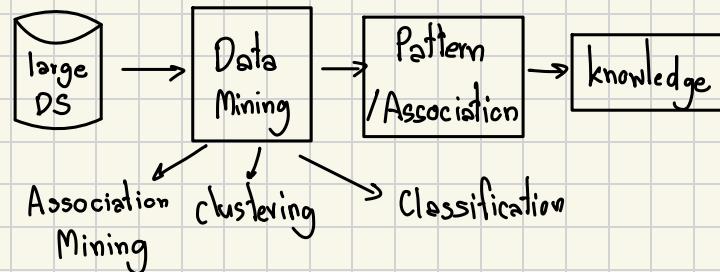
Data Mining overview

Knowledge Discovery in large Database (KDD)

คือกระบวนการที่กระทำกับข้อมูลวิเคราะห์นำมายield คุณภาพ pattern

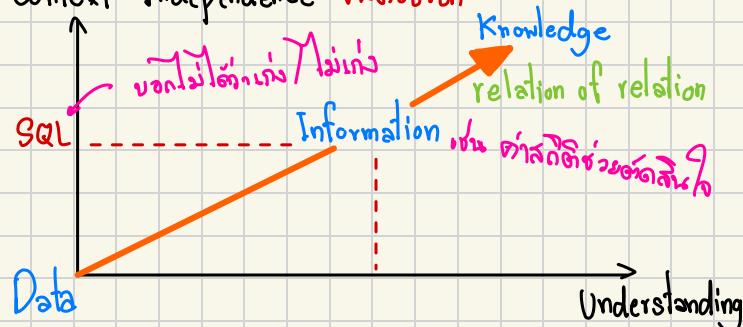
และความซับซ้อนซึ่งอยู่ในชุดข้อมูลนั้น ซึ่งไม่สามารถทำได้

ด้วยมือ โดยมุ่งเน้นกำกับในเรื่องการหานี้ของข้อมูลถึงการ
ตัดสินใจ ความซับซ้อนของข้อมูล จึงต้องใช้ technique (ทักษะ)
ตัดสินใจที่มีความซับซ้อน ความซับซ้อนของข้อมูลนั้นๆ จึงต้องใช้ technique (ทักษะ)



Pattern = เทคนิคที่เก็บข้อมูลเพื่อสร้างรูปแบบที่สามารถอ่านออกมายield (knowledge representation)

Context Independence ขึ้นต่อสืบ



Inductive Reasoning

- การนำข้อมูลมาสรุปผ่านขั้นตอน การนำข้อมูลมาหาความเหมือนกันในชุดข้อมูลนั้นๆ หรือกินความทุกทางโดยอ้อม ตามไปตามมา หาความเหมือนกันที่เกิดขึ้น ชุดข้อมูลเดียวกันนี้ แต่ต้องมีความต่างๆ กัน จึงสามารถเก็บข้อมูลนั้นๆ ให้เป็นข้อมูลเดียวกันได้
- การนำข้อมูลเดียวกัน ให้ความเหมือน
- การนำข้อมูลเดียวกัน ให้ความไม่เหมือน
- Inductive หาอะไรจริงๆ ไม่ได้ จำเป็นต้องใช้ logic

Perception (สัญชาตา, การรับรู้) คือกระบวนการที่จะนำข้อมูลเข้าสู่กัน แล้ว รับๆ ๆ ๆ โอบตัวเอง ณ. ที่สัมผัสได้ ที่เก็บสิ่งที่เราสนใจไว้และสามารถดึง

Perception ≠ concept concept คือการนำความรู้แล้ว visualise ให้ได้

Data Warehouse

คือส่วนตัวของระบบฐานข้อมูลสำหรับคุณภาพ เช่น กระบวนการผลิตเบลเยียมรัฐโรมันห้องแม่ค้าต่อตัวกันและกันเพื่อเตรียมการขายต่อในร้าน ซึ่งต้องการใช้ ETL - Extract, Transform, Load

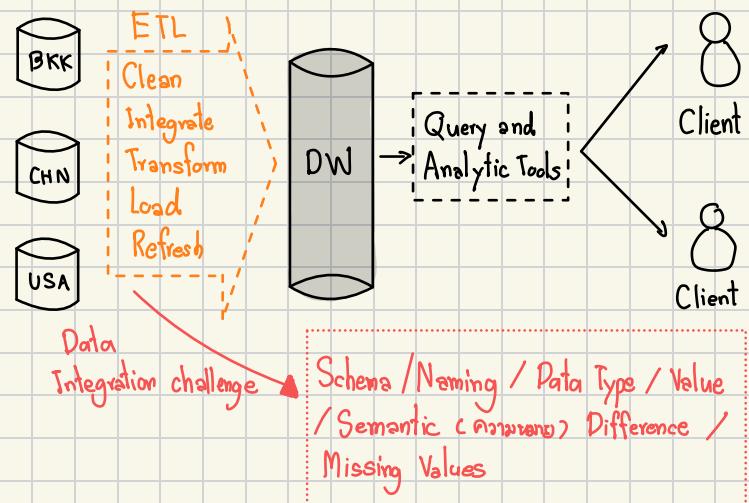
ประกอบด้วย การเก็บรวบรวม, การปรับปรุงข้อมูล, online analytical processing (OLAP)

OLAP เป็นเทคโนโลยีในการวิเคราะห์ข้อมูลที่มีประสิทธิภาพ รวมถึง การรวมตัว (consolidation) และการรวมกัน

DB + SQL	DW + OLAP
ดำเนินการ operation รายวัน ข้อมูลคงที่ไม่เปลี่ยน R / W ไม่มีข้อมูล redundancy	ดำเนินการทางเดียว ใช้ช่องลูกรักษาตัวเดียว Read only มีข้อมูล redundancy

Redundancy lead to anomaly ถ้ากระทำการลบ ที่มี冗余 ก็จะลบตัวเดียว แต่ต้องลบตัวที่

Data Warehouse Architecture



Knowledge constraints

- ① มีสสาร (Nontrivial)
- ② มีความถูกต้อง ใช้ได้จริง (Valid)
- ③ ไม่เคยทราบมาก่อน (novel / previously unknown)
- ④ นำไปใช้ได้ประโยชน์ (potentially useful)
- ⑤ น่าสนใจ (Interesting)
- ⑥ สามารถทำความเข้าใจ (Understandable)

Learning Category

- Supervised ผู้สอนสั่งที่ต้องรับจากอาจารย์ท่านนั้นชื่อชั้นคลาสชื่อบุนเด็จกานต์
ประเวศตัวอย่างมาคาดคะเน
- Unsupervised กรณีที่ไม่มีสารบัญ หรือไม่ชัดเจนว่า หรือสิ่งที่สอนในไปตั้งแต่ไหน เน้นการภาระงานตัวของข้ออธิบาย
โดยนิยามความต่างของข้อมูล

Major Data Mining Tasks

- ① Classification - กำหนด instance class
- ② Association - A B C เกิดพร้อมกัน
- ③ Clustering - ตัดหากลุ่มข้อมูลที่มีความคล้ายคลึงกัน
- ④ Outlier Mining - หาจุดข้างนอกเรื่องของการฝึก
- ⑤ Trend and Evolution Analysis - time-series mining
- ⑥ Mining path Traversal Patterns - ตัดหากรุํปแบบการเดินทางท่องเที่ยว
เช่นเดินทาง Web access log



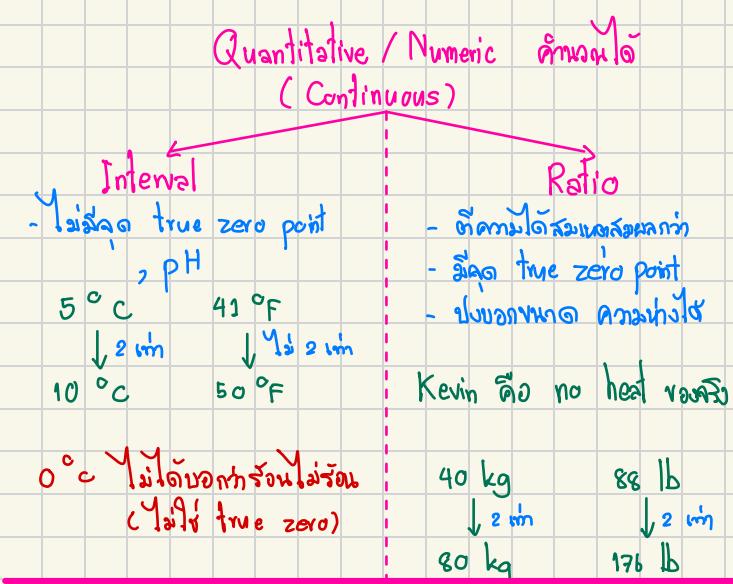
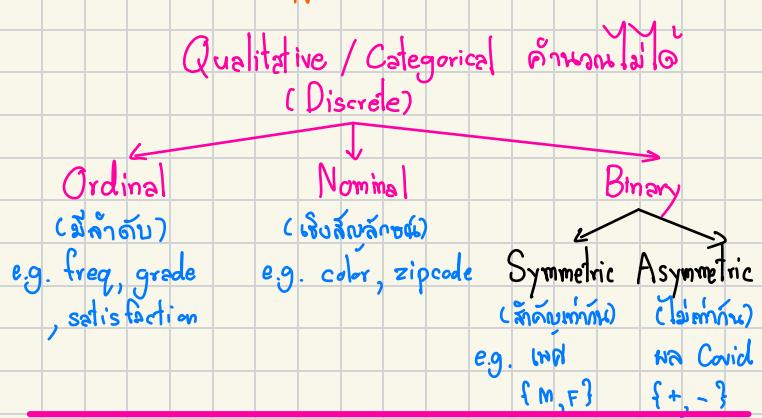
Data Preprocessing

Data Object

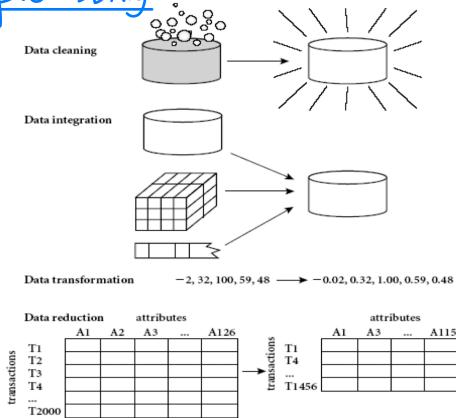
- មុនពេលក្រុករក្សាយ data object ទូទៅ
- យោងជាលំ entity ដែល
គ្មានចំណែកអាជីវកម្មខាងក្រោម : ផែករៀងនា, ភាគរូប, រាយកិច្ច
- បងកំកើងឱ្យការ samples, examples, instances, data points, object, tuples
- ក្នុងបីបាយចំណួល set of attributes

1 point = មុនពេលក្រុក
7 និង "

Attribute Data Types



Data Preprocessing



Data Cleansing

Missing Data

- សារព័ត៌មាន
សំណង់ បងកំណត់ថ្មី, ទៅត្រូវកិច្ច
សំណង់
- 1) ប្រកួតចិត្ត ពេលបងកំណត់ថ្មីដោយការ
ចែករាយប៉ុណ្ណោះទៅក្នុងក្រុង
 - 2) ឯក attribute ដើម្បីតាមរយៈប្រព័ន្ធប្រព័ន្ធ
 - 3) ឯកតាមលក្ខណៈ ឬ -1, 999, 999
 - 4) ប្រារិនតាមទំនួរ

- ពេលបងកំណត់ថ្មី ឲ្យត្រូវកិច្ចបាន
ទៅដោតស្អែកការណ៍ ឲ្យបាននិយោបាយ

- យោងតាមការ attribute ទិន្នន័យការណ៍
ដែល អូនុវិធីត្រូវនិងក្នុងការណ៍

9. model ការងារ

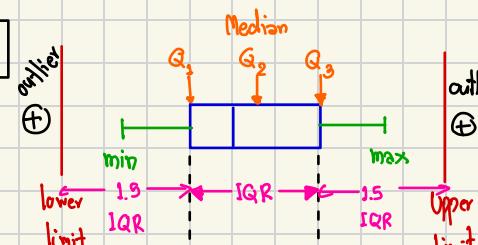
ចំណុចតាមការណ៍ (Central Tendency)
ដែលការងារ bell curve ទៅនឹងបងកំណត់ថ្មី
ដោយតាមការងារ

- Nominal មុនពេល mode
- Ordinal មុនពេល median
- Interval / Ratio ក្នុងថ្មី មុនពេល mean
- Interval / Ratio ក្នុងថ្មី មុនពេល median

ការការងារការងារ (Spread / Dispersion)
ដោយការងារ ផ្លូវការទៅគ្រប់គ្រងទំនាក់ទំនង

Interquartile Range (IQR)

$$IQR = Q_3 - Q_1$$



Step ការគាំរាយ

- ① ឲ្យបងកំណត់ថ្មី → ឲ្យការ
- ② ឲ្យ median
- ③ ដំឡើងចំណុចតាមការណ៍ median
ឲ្យបងកំណត់ថ្មី Q_1, Q_3
- ④ ឲ្យ IQR ;
- ⑤ ឲ្យ lower / upper limit
ឲ្យ $IQR \times 1.5$

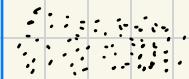
1	2	5	6	7	9	12	15	18	19	27
1	2	5	6	7	9	12	15	18	19	27

$(1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27)$

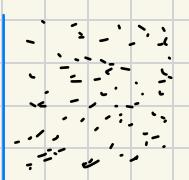
$$Q_1 = 5 \quad Q_3 = 18$$

$$\begin{aligned} IQR &= 18 - 5 = 13 \\ \text{Upper} &= Q_3 + 1.5 \text{IQR} = 37.5 \\ \text{Lower} &= Q_1 - 1.5 \text{IQR} = -14.5 \end{aligned}$$

Positively



Negatively



Correlated data

Uncorrelated data (Smear)

Data Transformation

- หลัก model ใน sklearn กำหนดให้ต้องมีลักษณะเดียวกันกับ input ของ model การแยกแยะปัจจัย ยกตัว tree-based model

Standization / Scaling

- แปลงข้อมูลเป็นมาตรฐานเดียวกันเพื่อความซึ่งรู้สึกที่ต่างกัน ที่ -1 ถึง 1
- ใช้กับ algorithm เช่น distance-based หรือ SVM KNN

Data Scaling : Sigmoidal แปลงในช่วง [-1, 1] ซึ่งมีประโยชน์ต่อการเรียนรู้ของเครื่องจักรการจำแนกประเภท

$$y' = \frac{1 - e^{-\alpha}}{1 + e^{-\alpha}} \quad \text{โดย} \quad \alpha = \frac{y - \bar{x}}{\sigma}$$

Normalization

- แปลงข้อมูลที่มีค่าตัวอย่างสูงให้ต่ำลง สำหรับการเรียนรู้ของเครื่องจักร (Gaussian Distribution)
- บ่งบอกการกระจายตัวของ Histogram และ Box plot

StandardScaler() $\rightarrow x' = \frac{x - \bar{x}}{\sigma}$ (z-score)

Skewed Distribution $\rightarrow x' = \log x$ (log transform / PowerTransformer())

MinMaxScaler() $\rightarrow \frac{v - \min}{\max - \min} = \frac{v' - \min'}{\max' - \min'}$

ตัด outlier แยกกันเมื่ออยู่ด้วย

Standard Scaler แยกกันข้อมูลที่ outlier ไม่ถูก Power Transformer แยกกันข้อมูลที่เปลี่ยนแปลงการคำนวณการแยกแยะปัจจัย

Data Type Conversion

Label encoding \rightarrow แปลง categorical เป็นตัวเลข แต่จะเก็บความพิเศษของค่า เช่น เดือนที่มีค่าเดียวกันที่ต้องการ

One hot encoding \rightarrow แปลง categorical เป็น binary n bits (Sparse matrix) แต่จะเก็บความพิเศษของค่า เช่น เดือนที่ต้องการ

Binning \rightarrow แปลง numeric เป็น categorical โดยเรียงข้อมูลลง

Equi-width แบ่งเท่ากัน ที่ แบ่งช่วงกว้าง 10 [-, 10) [10, 20) [20, +)

Alternative Equi-width ช่วงกว้าง = $\frac{\max - \min}{\text{จำนวนช่องแบ่ง}}$

Equi-depth / frequency ใช้ค่าเฉลี่ยของ N ที่ ทั้ง 3

1) 0, 4, 12 [-, 14) เดือน

2) 16, 16, 18 [14, 21) รุ่นรถ

3) 24, 26, 28 [21, +) ผู้หญิง

Equi-depth = 4

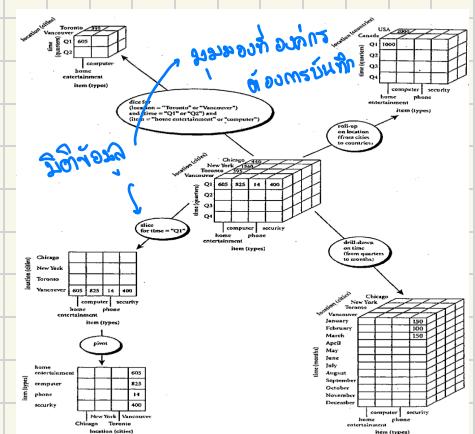


Data Reduction

ลดข้อมูลโดยใช้การรวม ที่เก็บไว้ใน Data Cube in data warehouse

ลดข้อมูลโดยเก็บข้อมูลรวมของรายเดือน ที่เก็บไว้ในโหนดเดียว (Data Cube)

OLAP



- Dimensionality Reduction / Feature Selection

การลด column ใด feature ที่ไม่เกี่ยวข้องกับการ train (irrelevant)

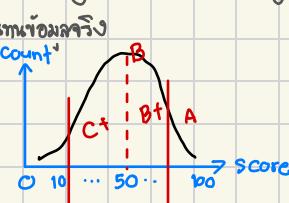
- ลบ Primary key (ID)
 - ลบ column ที่ไม่ใช้สำหรับ train
 - ลบ column ที่มีค่าหาย > 50%
 - ลบ column ที่ไม่ต่อเนื่อง เผื่องจาก column ที่ลบแล้วใน section นี้เดียวๆ
- * ปัจจุบันลด column ได้มากที่สุด = 2^{n-features}

- Principal Component Analysis (PCA)

- รูปแบบ eigenvalue และ eigenvector
- principal component ให้มาการวนิยมและปรับเปลี่ยนช่วงมาด้านหลัง column เช่น BMI

- Numerosity Reduction

- แบ่งค่าข้อมูลลงค่าซึ่งแทนที่จุดเดียว
- ใช้ Histogram หรือ Clustering เผื่องการกรุ๊ป化 แล้วเรียงตัวตาม

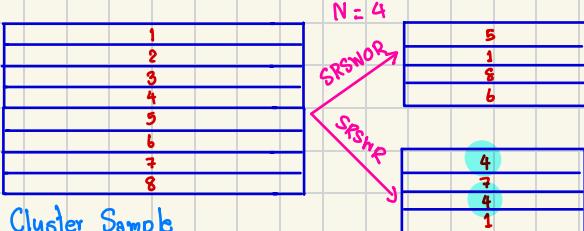


Class GPA = A - A
• 3.5, 3.7 = B+

Instance Selection (Sampling)

Simple Random Sample → สุ่มแบบไม่มีการกลับคืน
Without Replacement (SRSWOR) ไม่ได้แกว่งซ้ำ

Simple Random Sample → สุ่มแบบมีการกลับคืน
With Replacement (SRSWR) ห้องตัวอิสระ (uniform)



Cluster Samples

กำหนดจุดศูนย์กลาง 3 จุด และส่วนของแต่ละจุด

T_1, T_2, \dots, T_{100} $m=2 \rightarrow T_{201}, T_{202}, \dots, T_{300}$

$T_{101}, T_{102}, \dots, T_{200}$ $T_{701}, T_{702}, \dots, T_{800}$

Stratified sampling

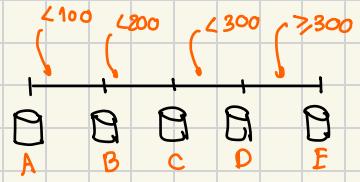
- Proportional allocation กำหนดอัตราส่วน
- Equal sample sizes กำหนดเมือง class ให้เท่ากัน

young	$y : m : s$	young
young		young
young		middle-aged
middle-aged	$2:2:2$	senior
senior		young
senior		young

Discretization

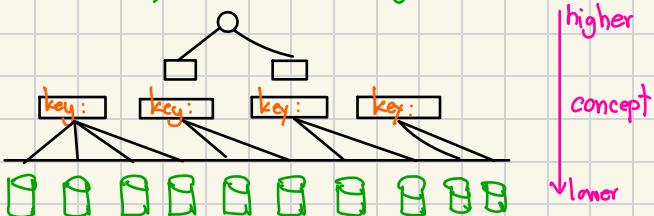
- หักห้ามจัดกลุ่มที่ต้องไม่ต่อเนื่อง

- ใช้สี binning



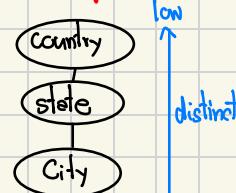
Concept Hierarchy

- ลักษณะ Categorical ต้องการสร้างลำดับชั้นในตัวเอง โดยมีความคล้ายคลึงมากที่สุด ให้เป็น concept ระดับสูงๆ
- Higher concept มีความนิ่นพื้นที่ไปมากกว่า lower concept
- ตัวอย่าง B-tree, B+ tree เพื่อ indexing direct access search

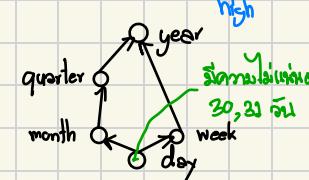


Schema Hierarchy ความสัมพันธ์ระหว่าง attribute ในฐานข้อมูล

Total order → ต้อง distinct value
ต้องมีรายละเอียดที่ต่างกัน

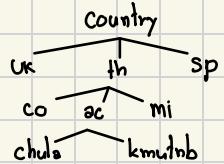
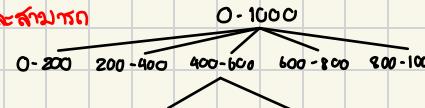


Partial order → จัดลำดับโดยไม่ต้องต่อเนื่อง
(สเกลลาร์ lattice)



Set-grouping hierarchy

แบบ attribute ของปีนี้จะ ไม่สามารถ
หักห้ามค่าเดียวได้



e.g., P1 = retail price of X

P2 = actual cost of X

lowProfitMargin(X) ← price(X, P1)

and cost(X, P2)

and (P1 - P2) < \$50

Rule-based hierarchy

การกำหนดค่าตัวเขียว ถ้าอิฐก้อนนี้

หักห้ามค่าเดียว

หากก้อนดินนี้มีบ่อวัว หักห้ามค่าเดียว

ยอดขายต่ำ < \$50 หักห้ามค่าเดียว

Data Preparation

1. รีบูตระบบ ตรวจสอบข้อผิดพลาด

2. เลือก attribute ที่เหมาะสม ลดจำนวนค่าที่ต้องคำนึงถึงเชิงคุณภาพ attribute

3. เลือก attribute ที่มีค่าเดียว, คำนึงถึงค่า attribute ที่ต้องการ

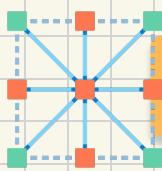
4. เลือก attribute

4.1 ลบ attribute ที่ไม่คุณภาพ (e.g. missing value > 50%)

4.2 ลบ missing value

4.3 ตรวจสอบ categorical feature เพื่อแปลงข้อมูล

4.4 ตรวจสอบ numeric เพื่อกำจัด outlier, หรือ normalization



Association Mining

Association : $X \rightarrow Y$ X เกิดแล้ว Y เกิดตาม \rightarrow กฎสังเกตุพิสูจน์

ตัวแปรความถี่สูง = ค่าอัตรา = $\frac{\# \text{ occurrence}}{\text{total}}$

ตัวแปรให้โอกาสเป็นไปได้, Pr = Prob Given = $P(Y|X)$

ความถี่สูงที่สุดของ Y เมื่อ X เกิดแล้ว

Association mining คือหาความสัมพันธ์ที่มีลักษณะว่า item set ใดๆ ไปร่วมกับ item set อื่นๆ ของ transaction ที่มี item ที่เป็นไปได้ (antecedent) ไปร่วม (consequent)

{Cheese, Bread} \rightarrow Bread

Application

- วิเคราะห์การซื้อสินค้า (Market Basket Analysis : MBA) เช่นใน Clustering อย่างนี้
- หาผลลัพธ์ของรายการซื้อขายที่มีใน Transaction หนึ่ง ที่เก็บ Transaction ณ จุด (point-of-sale) ผ่านมาทางบริการช้อปปิ้งเดลิเวอรี่
- MBA + วิเคราะห์กลุ่มเส้นทางที่ลูกค้าซื้อ หรือแพ็คเกจ bundle

ขั้นตอน Association mining

- Transaction T ต้องประกอบด้วย item $T \subseteq I$
- บริษัทฯ D ตรวจสอบ T หากเป็นไปได้
- T มี item สองตัว $x \times y$ ที่ $x \subseteq T$
- Association Rule คือการอุปนัย $x \rightarrow y$ โดย $x \subseteq I$, $y \subseteq I$ แล้ว $x \cap y = \emptyset$

e.g. $T_3 = \{I_4, I_{10}, I_{22}\}$

จำนวน itemset ย่อย = $2^3 - 1 = 7$

1. $I_4 \rightarrow I_{10} I_{22}$
2. $I_{10} \rightarrow I_4 I_{22}$
3. $I_{22} \rightarrow I_4 I_{10}$
4. $I_4 I_{10} \rightarrow I_{22}$
5. $I_4 I_{22} \rightarrow I_{10}$
6. $I_{10} I_{22} \rightarrow I_4$
7. \emptyset

Rule Basic Measure

- Support ของ item set A ให้ครองทั้ง item set B ใน T ต้องคำนึงถึง

$$S(A \rightarrow B) = P(A \cup B) = \frac{\# \text{ Transaction}(A \cup B)}{\# \text{ Trans ID}}$$

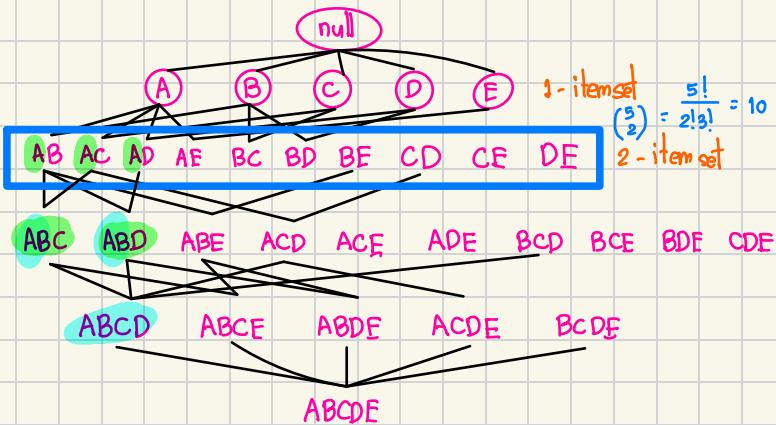
- Confidence ของชุด itemset T หากต้องมี itemset A รวมกับ itemset B

$$C(A \rightarrow B) = P(B|A) = \frac{\# \text{ Transaction}(A \cup B)}{\# \text{ Trans}(A)}$$

Itemset Lattice

- ผลลัพธ์ item set ที่มีปัจจัย $= 2^n - 1$

- ต่อเนื่องของ item แบบ level wise



Apriori Principle

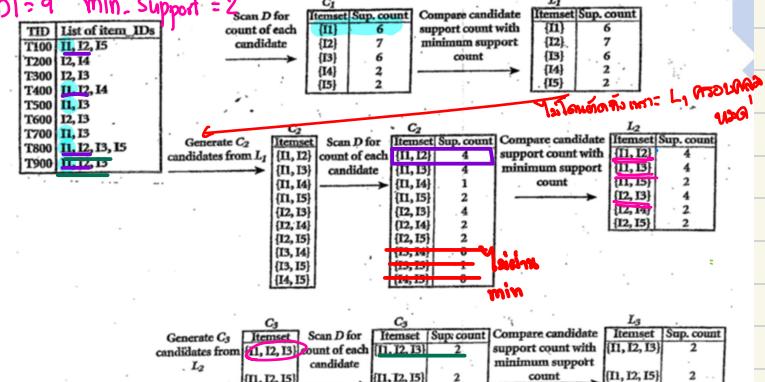
- ตัวแปร itemset ที่มีตัวอักษร frequent item ตัวเดียว item set ที่มีตัวอักษร เหลืออยู่น้อยกว่า min support
- ผลลัพธ์ของการ Anti-monotone property \rightarrow กรณี $P(x) < \text{min support}$ เลย superset x ของ $P(x \cup A) < \text{min support}$ ไม่อาจมีตัวอักษรมากกว่า min support

2 key step of Association Mining

1. ค้นหา itemset ที่มีตัวอักษร min support

2. หากฎความสัมพันธ์ของ itemset ที่มีตัวอักษร min confidence

$|I| = 9$ $\text{min. support} = 2$



L_3 ตัวอักษรที่มี min support: L_1 ตัวอักษรที่มี min support

$Hemset C_3 = L_2 \cap L_2 = \{ \{I_1, I_2, I_3\}, \{I_1, I_2, I_5\}, \{I_1, I_3, I_5\}, \{I_2, I_3, I_5\} \}$

$\{I_1, I_2, I_3\} = \{ \{I_1, I_2\}, \{I_1, I_3\}, \{I_2, I_3\} \}$ ตัวอักษรที่มี min support

ตัวอักษรที่มี min support จาก L_2

$I_1 \rightarrow I_2$

$I_2 \rightarrow I_1$

$I_1 \rightarrow I_3$

$I_3 \rightarrow I_1$

\vdots

$I_2 \rightarrow I_5$

$I_5 \rightarrow I_2$

confidence

$I_1 \rightarrow I_5$

$I_5 \rightarrow I_1$

$\{I_3, I_5\} \notin L_2$

$$C(I_1 \rightarrow I_2) = P(I_2 | I_1) = \frac{\# \text{ Trans}(I_1 \cup I_2)}{\# \text{ Trans}(I_1)} = \frac{4}{6} < 70\%$$

$I_3 \in L_2$

$I_5 \in L_2$

$I_1 \in L_2$

$I_2 \in L_2$

$I_3 \in L_2$

$I_5 \in L_2$

$I_1 \in L_2$

$I_2 \in L_2$

$I_3 \in L_2$

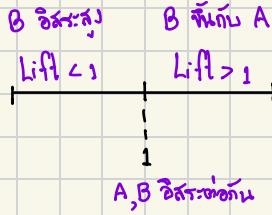
Dependent Framework

- การเกิดเหตุการณ์ A นำไปสู่เหตุการณ์ B (imply) การเกิด B
กรณีที่เกิดค่ามากกว่า Strong Association Rules มากกว่าเดิม
- การอธิบายความพึ่งพา A ให้ด้วยความน่าจะเป็นของ B ในเมื่อ A ไม่เกิดขึ้นแล้ว
- ความน่าจะเป็นของ A และ B ความน่าจะเป็นของ B ในเมื่อ A ไม่เกิดขึ้นแล้ว

Correlation / Lift / Interest

$$\text{Lift}(A \rightarrow B) = \frac{P(B|A)}{P(B)}$$

$$= \frac{P(A \cup B)}{P(A)P(B)}$$



X	1	1	1	1	0	0	0	0
Y	1	1	0	0	0	0	0	0
Z	0	1	1	1	1	1	1	1

$S(X \rightarrow Y) = \frac{\# \text{Trans}(X \cup Y)}{\# \text{Trans}(Y)} = \frac{2}{8}$

$$C(X \rightarrow Y) = \frac{\# \text{Trans}(X \cup Y)}{\# \text{Trans}(X)} = \frac{2}{4}$$

$$\text{Lift}(X \rightarrow Y) = \frac{C(X \rightarrow Y)}{P(Y)} = \frac{2/4}{2/8} = 2$$

Clustering



Natural Grouping

- ใช้เป็นเครื่องมือในการน้ำกรากษาข้อมูลของข้อมูล
หรือการทำให้ data preprocessing เพื่อหาน outlier
- การจัดกลุ่มขึ้นอยู่กับเกณฑ์ที่ใช้ของข้อมูล
- แบ่งเป็นกลุ่มที่สีความคล้ายกัน ('High intra-class similarity')
และไม่คล้าย (Low intra-class similarity)

กิจกรรมการ Clustering ตอนนี้

1. กำหนดการ (เน้นขอตัวอย่างคุณลักษณะของ → DNA, text)
2. กำหนดหน่วยวัดระดับห่าง (ค่าความใกล้เคียง) → Euclidean
3. กำหนดวัสดุก่อสร้างการจัดกลุ่มที่จะใช้ → K-Mean

Minkowski Distance

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

$q = 1$ คือ Manhattan distance

$q = 2$ คือ Euclidean distance

คุณสมบัติของห้องวัดระยะห่าง

$$D(A, B) \geq 0$$

$$D(A, B) = D(B, A)$$

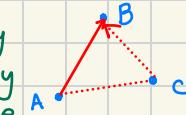
$$D(A, B) = 0 \text{ iff } A = B$$

$$D(A, B) \leq D(A, C) + D(C, B) \text{ Triangle Inequality}$$

Positivity

Symmetry

Reflexive



Binary Variable

- Symmetric → Simple Matching Coefficient

$$d(i, j) = \frac{b+c}{a+b+c+d} \rightarrow \text{พิจารณาค่าเดียวทั้งคู่}$$

- Asymmetric → Jaccard Coefficient

$$d(i, j) = \frac{b+c}{a+b+c}$$

$d(i, j) = 0$ คือเมื่อถูก attribute

$d(i, j) = p$ ถือถูกกันทุก attribute

Nominal Variable

- Simple Matching → นับจำนวนค่าที่ค่าในชื่อเท่ากัน

p คือจำนวนที่ไม่เท่ากัน

$$d(i, j) = \frac{p-m}{p}; p = a+b+c+d$$

- แทนตัวถ้าที่เป็นไปได้ m ค่า ตัวต่อไป binary m bits
(One hot encoding)

$$\begin{matrix} m=5 & 0 & 1 & 0 & 0 & 1 \\ & \downarrow & & \downarrow & & \end{matrix}$$

Ordinal Variable

- ใช้เป็นค่าระดับชั้น เส้นสูงต่ำ เป็น interval-scaled ตามเหตุผล "ค่าต่อตัว"
- แปลงค่าตัวเป็นร้อยละ f^{th} ของจุดต่ำที่ i ตรวจสอบ min-max scalar อยู่ในช่วง $[0, 1]$

$$\begin{aligned} Z_{if} - \min' &= \frac{v - \min}{\max' - \min'} & \begin{matrix} 1 & 2 & 3 & 4 & 5 \\ 0 & 0.25 & 0.5 & 0.75 & 1 \end{matrix} & \text{ต่อตัว} \\ Z_{if} - 0 &= \frac{r_{if} - 1}{M_f - 1} & \begin{matrix} r_{if} \\ M_f \end{matrix} & \end{aligned}$$

5 อยู่ใน max row column

Variables of Mixed Type

$$d(i, j) = \sum_{f=1}^P S_{ij}^{(f)} d_{ij}^{(f)}$$

$$S_{ij}^{(f)} = 0 \text{ ถ้า}$$

1. $x_{if} \neq x_{jf}$ สำหรับ x_{if}
2. f เป็น asymmetric binary และ $x_{if} = x_{jf} = 0$

$$S_{ij}^{(f)} = 1 \quad \text{กรณีที่} \quad x_{if} = x_{jf} \neq 0$$

Table 7.3 A sample data table containing variables of mixed type.

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)	log (test-2)
1	code-A	excellent	445	2.65
2	code-B	fair	22	1.34
3	code-C	good	164	2.21
4	code-A	excellent	1210	3.08

Dissimilarity Matrix of test-1

$$\begin{bmatrix} 0 & d(2,1) & d(3,1) & d(4,1) \\ d(2,1) & 0 & 1 & 0 \\ d(3,1) & 1 & 0 & 0.5 \\ d(4,1) & 0 & 0.5 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & d(2,2) & d(3,2) & d(4,2) \\ d(2,2) & 0 & 1 & 0 \\ d(3,2) & 1 & 0 & 0.5 \\ d(4,2) & 0 & 0.5 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & d(2,3) & d(3,3) & d(4,3) \\ d(2,3) & 1 & 0 & 0 \\ d(3,3) & 0 & 1 & 0.5 \\ d(4,3) & 0 & 0.5 & 0 \end{bmatrix}$$

$\text{ด้านซ้าย} = \text{ด้านขวา} = 2.65 - 1.34$

$2.21 - 1.34$

$$Z_{if} = \frac{2-1}{3-1} = 0.5$$

ratio-scale ต้อง normalize ด้วย min-max scalar ก่อนแล้วถึงจะสามารถ Variable mixed type

$$\begin{bmatrix} 0 & 1.31 & 0 & 0.44 & 0.87 & 0 \\ 1.31 & 0 & 0.44 & 0.87 & 0 & 0.43 \\ 0 & 0.44 & 0 & 1.74 & 0.87 & 0.43 \\ 0.43 & 0 & 1.74 & 0 & 0.87 & 0 \\ 0 & 0.87 & 0 & 0.43 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\substack{x \\ \text{max-min}}} \begin{bmatrix} 0 & 0.75 & 0 & 0.25 & 0.50 & 0 \\ 0.75 & 0 & 0.25 & 0.50 & 0 & 0.25 \\ 0 & 0.25 & 0 & 1.00 & 0.50 & 0 \\ 0.25 & 0 & 1.00 & 0 & 0.50 & 0 \\ 0 & 0.50 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & d(2,1) & d(3,1) & d(4,1) \\ d(2,1) & 0 & 1 & 0 \\ d(3,1) & 1 & 0 & 0.5 \\ d(4,1) & 0 & 0.5 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0.92 & 0 & 0.58 & 0.67 & 0 \\ 0.92 & 0 & 1 & 0.58 & 0.67 & 0 \\ 0 & 1 & 0 & 0.92 & 0.67 & 0 \\ 0.58 & 0 & 0.92 & 0 & 0.67 & 0 \\ 0 & 0.67 & 0 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.67 & 0 & 0 \end{bmatrix}$$

$\frac{1(1) + 1(1) + 1(0.75)}{3} = 0.92$

$\frac{1}{3}$ ตัว d ของ test 1 ตัว d ของ test 2

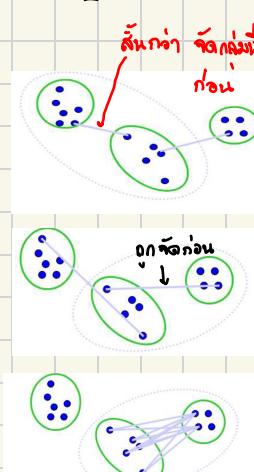
Cluster Similarity Criteria

- Single link เลือกสมการที่คล้ายสุดกันทุกคู่

- Complete link เลือกสมการที่ไม่คล้ายมากที่สุด

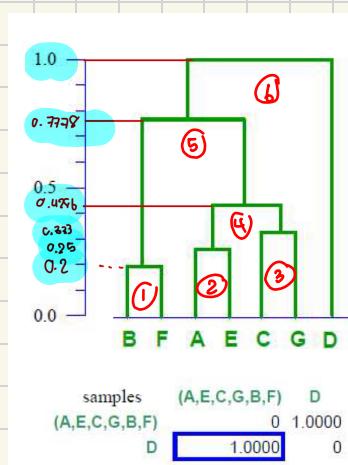
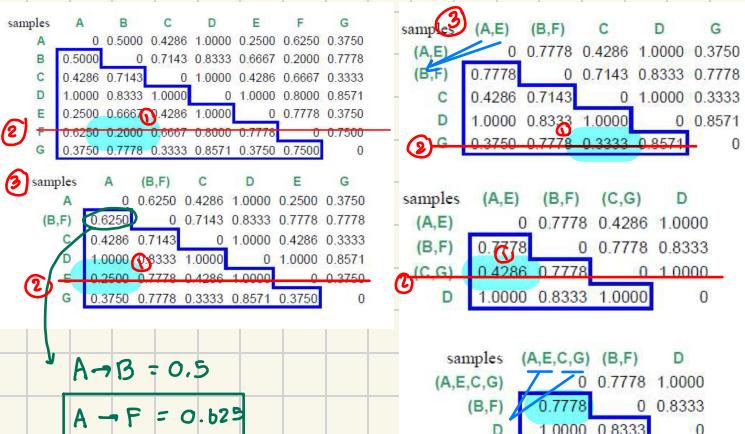
* outlier ไม่รวมอยู่ในกลุ่มก่อน

- Average Link หาราคาของทุกคู่ที่คล้าย



Hierarchical Clustering

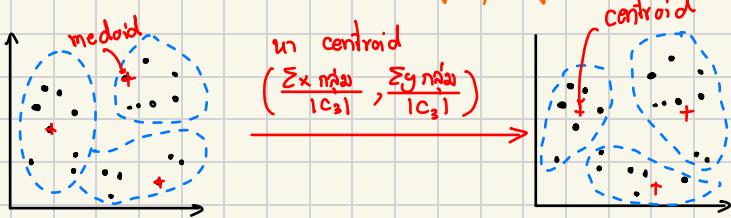
- ① เลือกค่าที่ต้องการจะหาน้ำยสูตร จากชั้นรวมกลุ่ม เดียวต่อไปก็ต้อง 1 群
- ② หาตัวแบบ dendrogram ตัวอย่างค่าจะหาน้ำยสูตรของกลุ่ม
- ③ ปรับปรุงตัวระยะห่าง ตัวกลุ่มนั้นมีมากกว่า 1 ถ้า ประยุกต์เก็บข้อมูลก็ต้อง เลือกตัวอย่างมากสูตร
- ④ ทำ ①-③ วนครับกากกุด



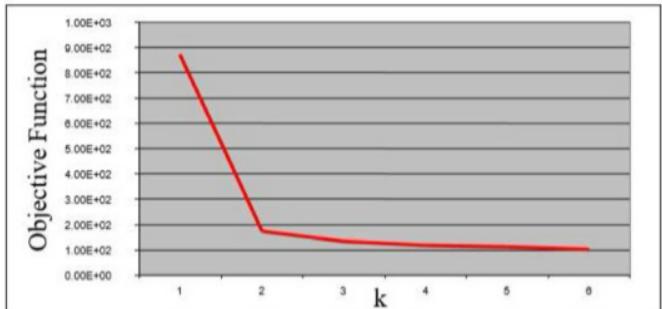
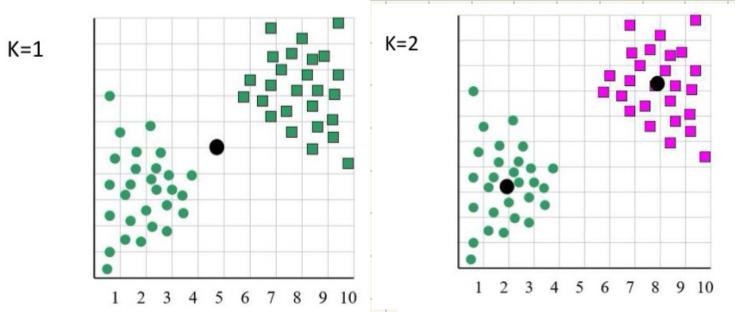
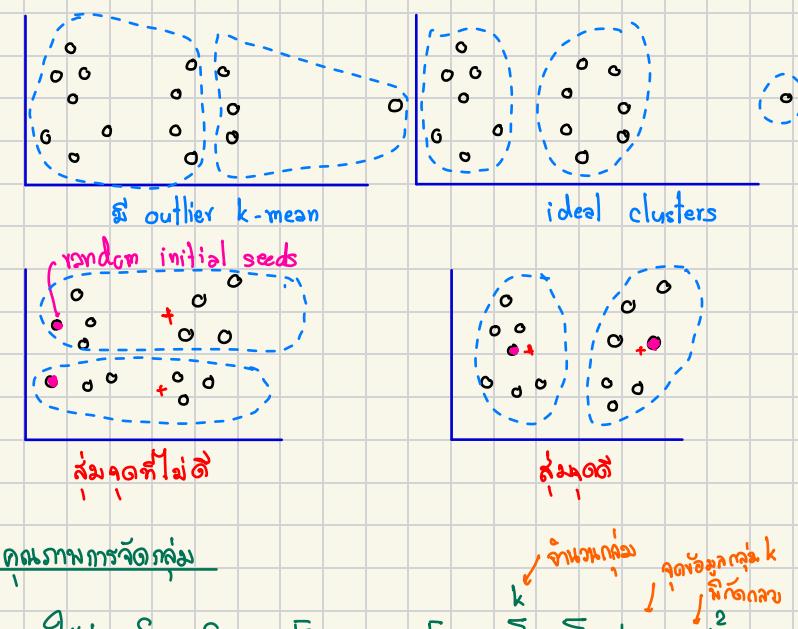
- ไม่ต้องระบุจำนวนกลุ่มที่ต้องการ
- เวลาการคำนวณไม่ต้องยืด → $O(n^2)$

Partitional Clustering (K-mean)

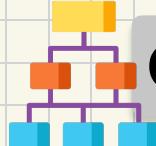
- Centroid เป็นจุดศูนย์กลางของกลุ่มที่ไม่ได้อยู่ในชั้นหัวหน้า
- Medoid เป็นจุดศูนย์กลางที่เหลือเชิง อยู่ในชั้นหัวหน้า



- ฟังก์ชันง่าย และเวลาการคำนวณ → $O(tkn)$
- กรณีต้อง mean ที่ต้อง categorical ไม่ได้
- ไม่สามารถจัดการกับ outlier ได้
- ไม่ต้องรู้รูปทรงของ cluster ในรูป non-convex



Classification



Dataset

- សម្រាប់សរុប (Training dataset) នឹងសម្រួលរាយរបស់វា
- សម្រាប់តាមសរុប (Testing dataset) នឹងវាសម្រាប់រាយរបស់វា
- សម្រាប់មុនពាណិជ្ជកម្ម (validation dataset) ប្រើបានដើម្បីសម្រាប់វា ដើម្បីរាយរបស់វា ដើម្បីរាយរបស់វា

Accuracy

$$\begin{aligned} \text{- Holdout} &= \frac{\sum_{i=1}^{|test_set|} S_i}{|test_set|}; \quad S_i = 1 \text{ តារាងអនុញ្ញាត} \\ &\qquad\qquad\qquad S_i = 0 \text{ តារាងអាយុយចិត្ត} \\ \text{- k-fold} &= \frac{\sum_{i=1}^k \sum_{j=1}^{|test_fold|} S_{ij}}{|train_dataset|}; \quad S_{ij} = 1 \text{ តារាងអនុញ្ញាត} \\ &\qquad\qquad\qquad S_{ij} = 0 \text{ តារាងអាយុយចិត្ត} \end{aligned}$$

Confusion Matrix

		model		Sensitivity (Recall)	Specificity
		Positive (+)	Negative (-)		
Positive (+)	True Positive (TP)	False Negative (FN)		$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$
	False Positive (FP)	True Negative (TN)			
Negative (-)	Precision	Negative Predictive Value		$\frac{TP}{TP+FP}$	$\frac{TN}{TN+FN}$

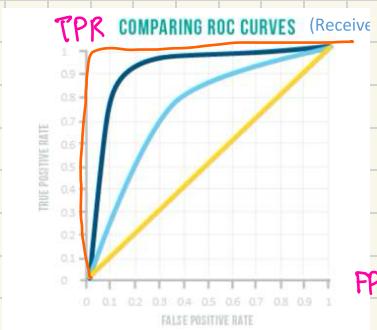
Precision = ទូទាត់អ្នករំភ័ណ៌ការងារ រាយរាយកំណត់

Recall (+) = ទូទាត់អ្នករំពីរីន + កំណងគេ រាយរាយកំណត់

Receiver Operating Characteristics Curve (ROC Curve)

- balance model តើនេះ TP ឬវិល FP នៅរបស់វា
- Area Under Curve (AUC) = 1 តើង model នឹងត្រូវស្វែងរក
- AUC = 0.5 តើង random guess

$$FPR = \frac{FP}{FP + TN}; \quad FPR \text{ or } \text{Specificity}$$



$$F\text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \text{TP}}{2 \text{TP} + \text{FP} + \text{FN}}$$

$$F\beta \text{ Score} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}$$

Deal class imbalance

- Oversampling - duplicate minority
- Under sampling - តើង majority class ធ្លាក់វិនិច្ឆ័យ តើង majority class រាយរាយ minority class