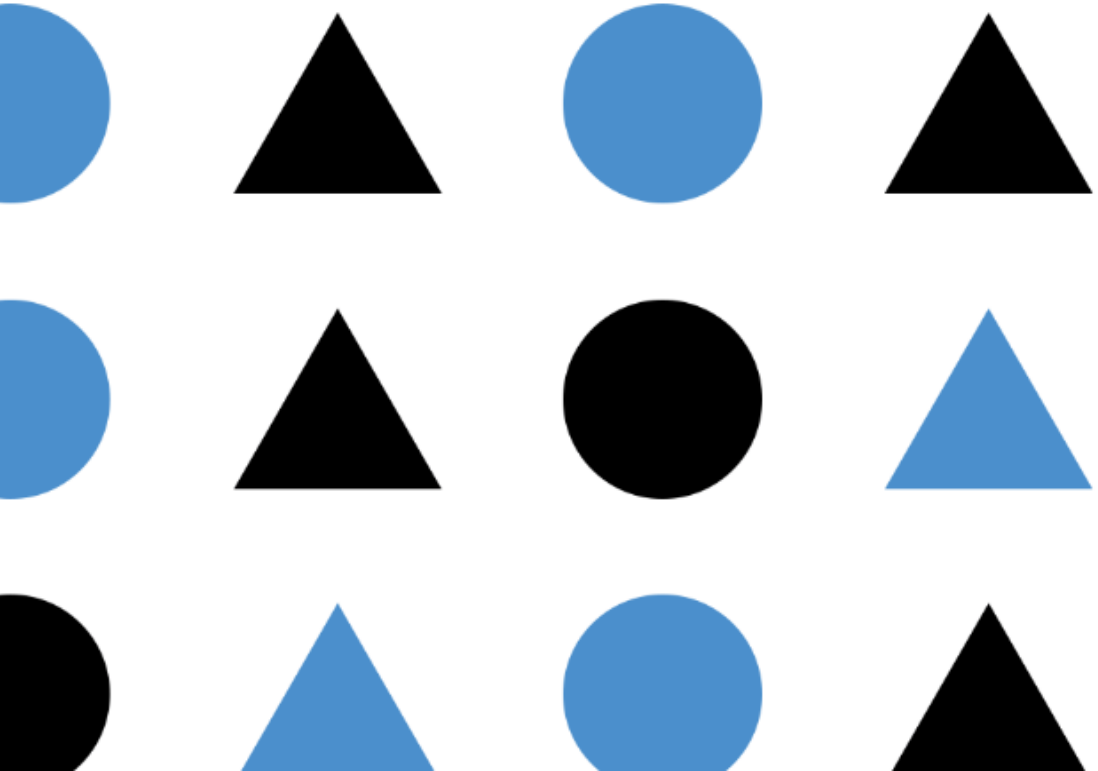


Unsupervised ML part II : Similarity and Dissimilarity

อ.ดร.ปัญญนัท อ้นพงษ์

ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร

aonpong_p@su.ac.th

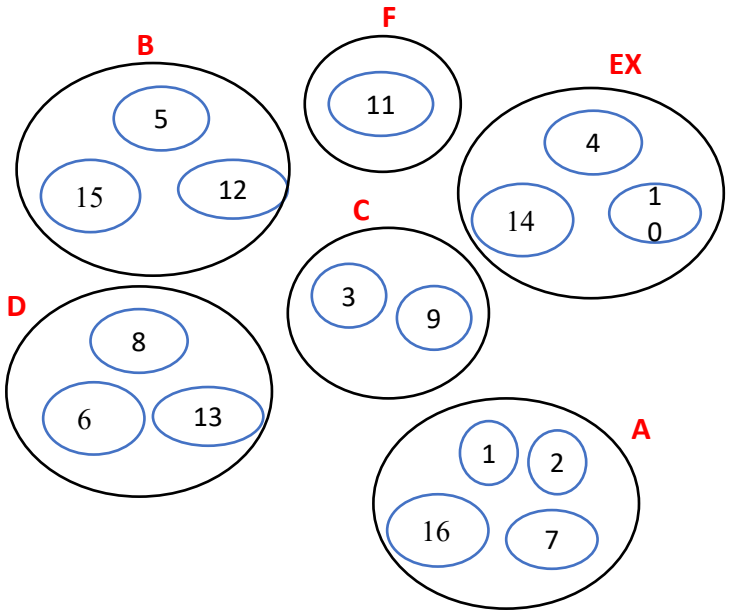
Outline

- **แนวคิดและแนวทางการประยุกต์ใช้ Clustering**
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- Definition of Clustering Problem
- Similarity / Dissimilarity measurement

แนวคิดและแนวทางการประยุกต์ใช้ Clustering

- Clustering ไม่ต้องรู้ class label ล่วงหน้า แต่ Supervised จำเป็นต้องรู้ class label ล่วงหน้า
- ให้นักศึกษาพิจารณาตัวอย่างต่อไปนี้
- จากตารางและภาพ เห็นได้ชัดเจนว่าเราใช้ข้อมูลจากตารางในการวาดแผนภาพ
  - หมายความว่าเรารู้ class label (Grade) จากข้อมูลที่มีเพื่อใช้ในการแบ่งประเภท (supervised classification)
- ลองจินตนาการว่าเราไม่รู้ class label (grade) ล่วงหน้า และเราต้องการจัดกลุ่มของข้อมูลเข้าด้วยกัน
  - สมมติเราต้องการแบ่งนักศึกษาออกเป็น 6 ระดับ ผลลัพธ์จะต่างจาก supervised classification อย่างไร

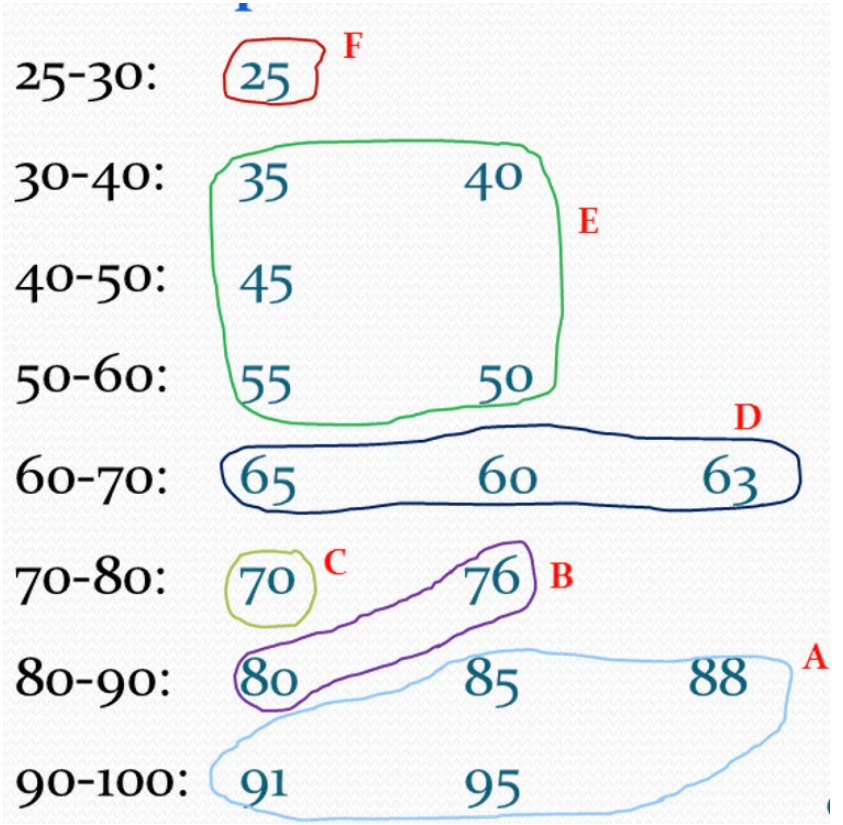
Roll No	Mark	Grade
1	80	A
2	70	A
3	55	C
4	91	EX
5	65	B
6	35	D
7	76	A
8	40	D
9	50	C
10	85	EX
11	25	F
12	60	B
13	45	D
14	95	EX
15	63	B
16	88	A



แนวคิดและแนวทางการประยุกต์ใช้ Clustering

- ผลลัพธ์ที่ได้จะแตกต่างจากการแบ่งแยกด้วยกระบวนการ Supervised Classification
- การแบ่งแยกแบบไม่มีการ predefine จะใช้การจัดกลุ่มโดยการหาความคล้ายคลึง (Similarity) ระหว่างข้อมูลตามที่สามารถพบได้บนปริภูมิ
- การจัดกลุ่มแบบนี้คือ clustering (Unsupervised)

Roll No	Mark	Grade
1	80	A
2	70	A
3	55	C
4	91	EX
5	65	B
6	35	D
7	76	A
8	40	D
9	50	C
10	85	EX
11	25	F
12	60	B
13	45	D
14	95	EX
15	63	B
16	88	A



แนวคิดและแนวทางการประยุกต์ใช้ Clustering

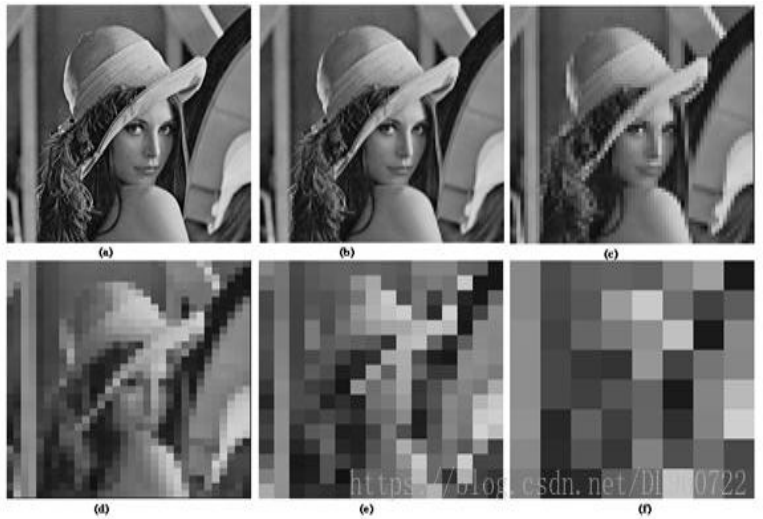
- เพื่อให้เห็นการประยุกต์ใช้ Clustering กับข้อมูลในชีวิตจริง ลองพิจารณาข้อมูลที่มีความซับซ้อนขึ้น

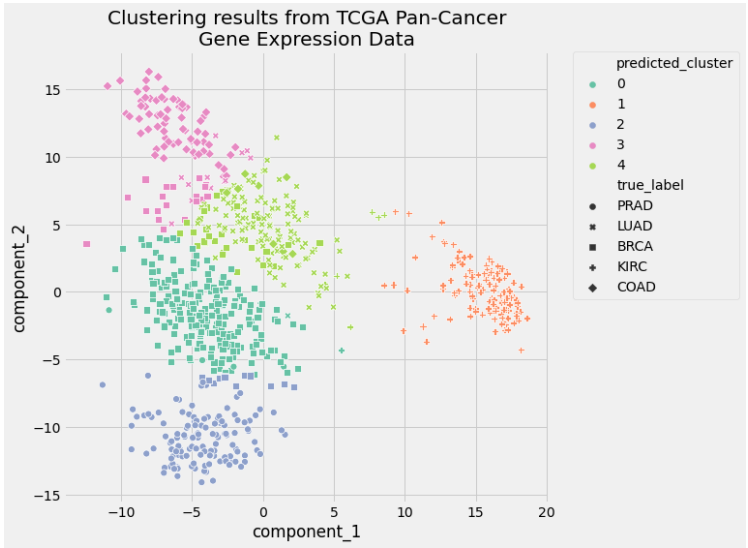
Martial Status	Age	Income	Education	Number of children
Single	35	25000	Under Graduate	3
Married	25	15000	Graduate	1
Single	40	20000	Under Graduate	0
Divorced	20	30000	Post-Graduate	0
Divorced	25	20000	Under Graduate	3
Married	60	70000	Graduate	0
Married	30	90000	Post-Graduate	0
Married	45	60000	Graduate	5
Divorced	50	80000	Under Graduate	2

- เมื่อข้อมูลมีหลาย Feature มากขึ้น เราสามารถทำ Clustering ข้อมูล 1 attribute หรือมากกว่า (กลุ่มของ attributes) ก็ได้ หากกลุ่มของ attribute นั้นมีความคล้ายคลึงกัน (high similarity)

แนวคิดและแนวทางการประยุกต์ใช้ Clustering

- มีการประยุกต์ใช้ Clustering กับงานจริงจำนวนมาก:
  - ใช้กับข้อมูลเสียงในการจำแนกเสียงพูด เพื่อจำแนกคลื่นเสียงจากเสียงพูดออกเป็น k ประเภท
  - ใช้สำหรับเลือกโทนสีบนอุปกรณ์แสดงผลกราฟิกแบบเก่าที่มีการจำกัดจำนวนสีและ Image Quantization (เรียกว่า Vector Quantization หรือ Image Segmentation)
  - การเรียกค้นเอกสาร
  - งานประเภทการเรียนรู้ของเครื่อง (machine learning) เป็นต้น



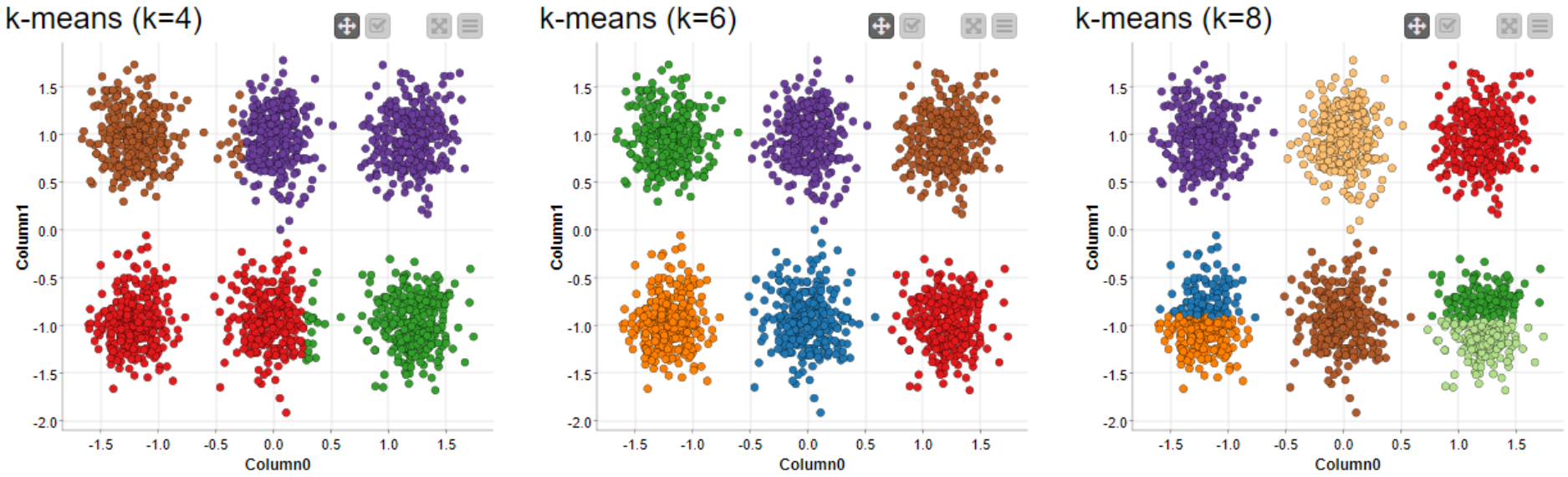


Outline

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- **ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง**
- Definition of Clustering Problem
- Similarity / Dissimilarity measurement

ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง

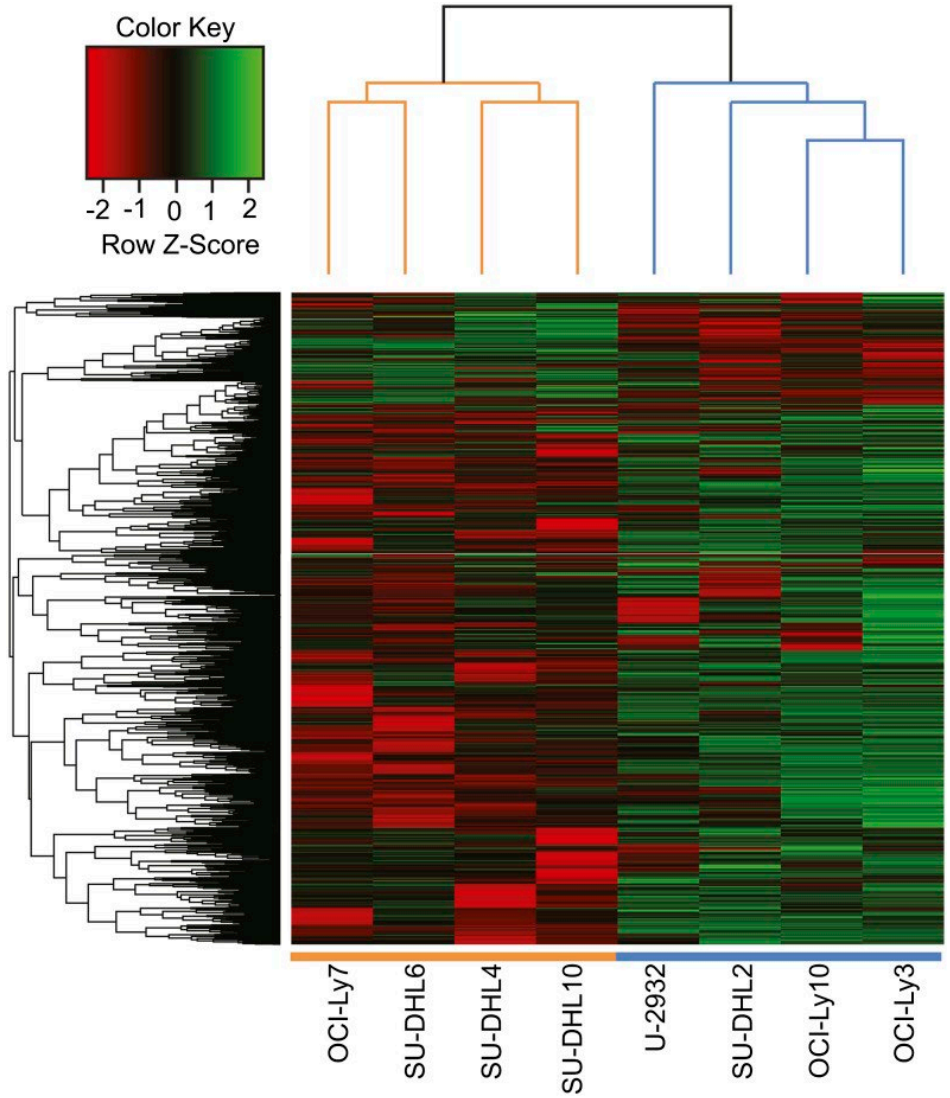
1. เราไม่ทราบจำนวนของคลัสเตอร์ที่เหมาะสมที่สุด
   - ไม่มีจำนวนคลัสเตอร์ที่เป็นคำตอบที่ถูกต้อง 100%
   - ในทางปฏิบัติ การทำการทดลองกับข้อมูลที่ใช้จริง ผู้ทำการศึกษาอาจพบว่ามีจำนวนที่เหมาะสมหลายจำนวน
   - ในข้อมูลขนาดใหญ่มาก การเลือกจำนวนที่เหมาะสมที่สุดที่แท้จริงนั้นไม่ใช่เรื่องง่าย



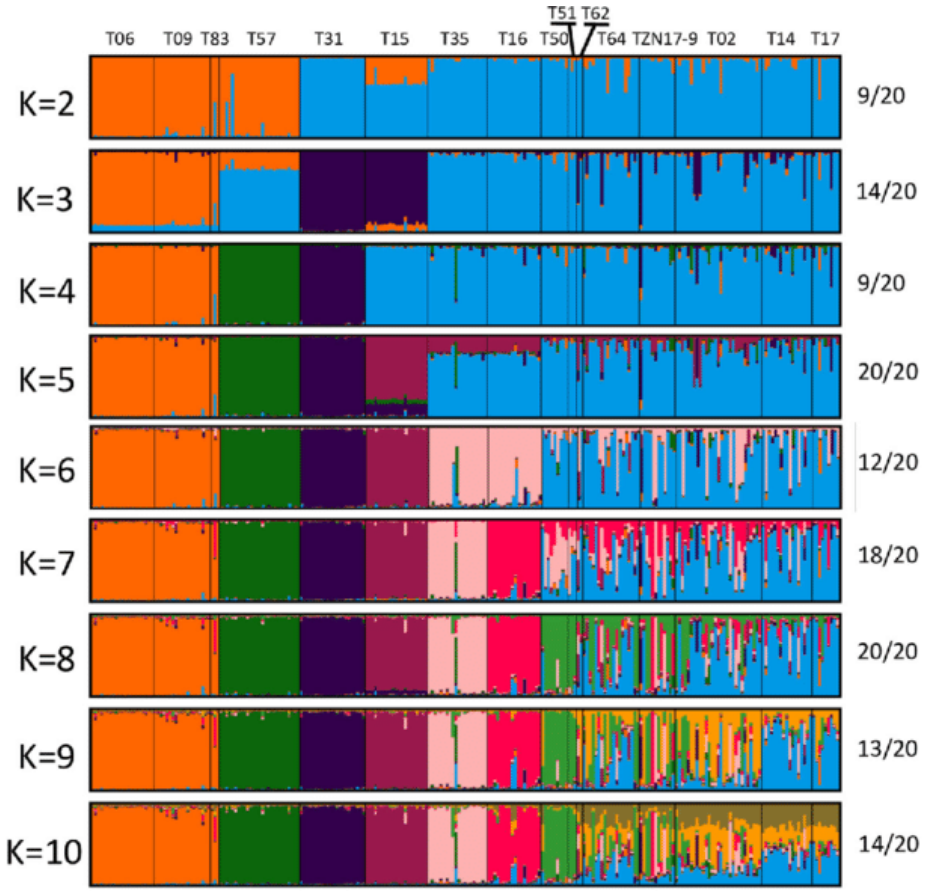
ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง

2. เราอาจไม่มีความรู้หรือความเข้าใจข้อมูลที่นำมาใช้ก็ได้

- ในทางปฏิบัติ หลายครั้งที่บริษัทหรือเจ้าของข้อมูลไม่ต้องการเปิดเผย Label ของข้อมูล เนื่องจากต้องการให้ข้อมูลนั้นเป็นความลับภายใน
- ผู้ศึกษาไม่อาจรู้ความหมายที่แท้จริงของข้อมูลได้ จึงจำเป็นต้องใช้ความรู้ทางสถิติเข้ามาจัดการ (ทั้งที่ไม่รู้ความหมายที่แท้จริง)
- หรือบางข้อมูล อาจไม่มีใครรู้หรือเข้าใจความหมายของข้อมูลเหล่านั้นมาก่อนเลยก็เป็นไปได้ (เราอาจเป็นคนแรกที่ทำการศึกษา)
- ปัญหานี้มักเป็นปัญหาของ clustering (เพราะถ้าไม่มีปัญหานี้ ในชีวิตจริงเรามักจะเลือกวิธีการ supervised classification)



ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง



3. การตีความหมายของแต่ละคลัสเตอร์เป็นเรื่องยาก

- การวิเคราะห์ด้วยวิธีการ Supervised Classification หรือการจำแนกแบบมีผู้สอน เราจะรู้อยู่แล้วว่าแต่ละ Class มี label เป็นอะไร เพราะในข้อมูลที่ใช้ฝึกฝนโมเดลจำเป็นต้องมี label ระบุอยู่แล้ว
- แต่ใน Unsupervised หรือ Clustering ส่วนใหญ่เราจะไม่ทราบ label แม้ว่าเราจะแยกข้อมูลเป็นกลุ่ม ๆ จากความคล้ายคลึง (Similarity) ออกจากกันได้ แต่การที่จะทราบความหมายที่แท้จริงของแต่ละกลุ่มนั้น จำเป็นต้องมีความรู้เกี่ยวกับข้อมูลมาก่อน หรือต้องมีการเข้าไปสำรวจข้อมูลภายในเท่านั้น
- ดังนั้น เมื่อกระบวนการ Clustering สำเร็จแล้ว ความหมายที่แท้จริงของคลัสเตอร์อาจไม่ชัดเจน

Outline

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- **Definition of Clustering Problem**
- Similarity / Dissimilarity measurement

Definition of Clustering Problem (แบบแฟนซี)

หนังสือเล่มหนึ่งได้กล่าวไว้ว่า

Definition : Clustering

Given a database $D = \{t_1, t_2, \ldots\ldots, t_n\}$ of n tuples, the clustering problem is to define a mapping $f : \mathrm{D} \rightarrow C$, where each $t_i \in D$ is assigned to one cluster $c_i \in C$. Here, $C = \{c_1, c_2, \ldots\ldots, c_k\}$ denotes a set of clusters.

- วิธีการแก้ปัญหา Clustering คือการกำหนดสูตร (mapping function) ในการวาดผังข้อมูล
- Mapping function ที่อยู่เบื้องหลังการวาดผังข้อมูลดังกล่าวคือการระบุว่าข้อมูลภายในคลัสเตอร์หนึ่งนั้นเหมือนข้อมูลภายในคลัสเตอร์นั้นมากกว่า และไม่เหมือนกับทูเพิลที่อยู่ภายนอก

Definition of Clustering Problem (แบบแฟนซี)

ดังนั้น mapping function จาก definition ที่ระบุไว้ อาจกล่าวได้อย่างชัดเจนว่า

$$f : \mathrm{D} \rightarrow \{c_1, c_2, \ldots\ldots, c_k\}$$

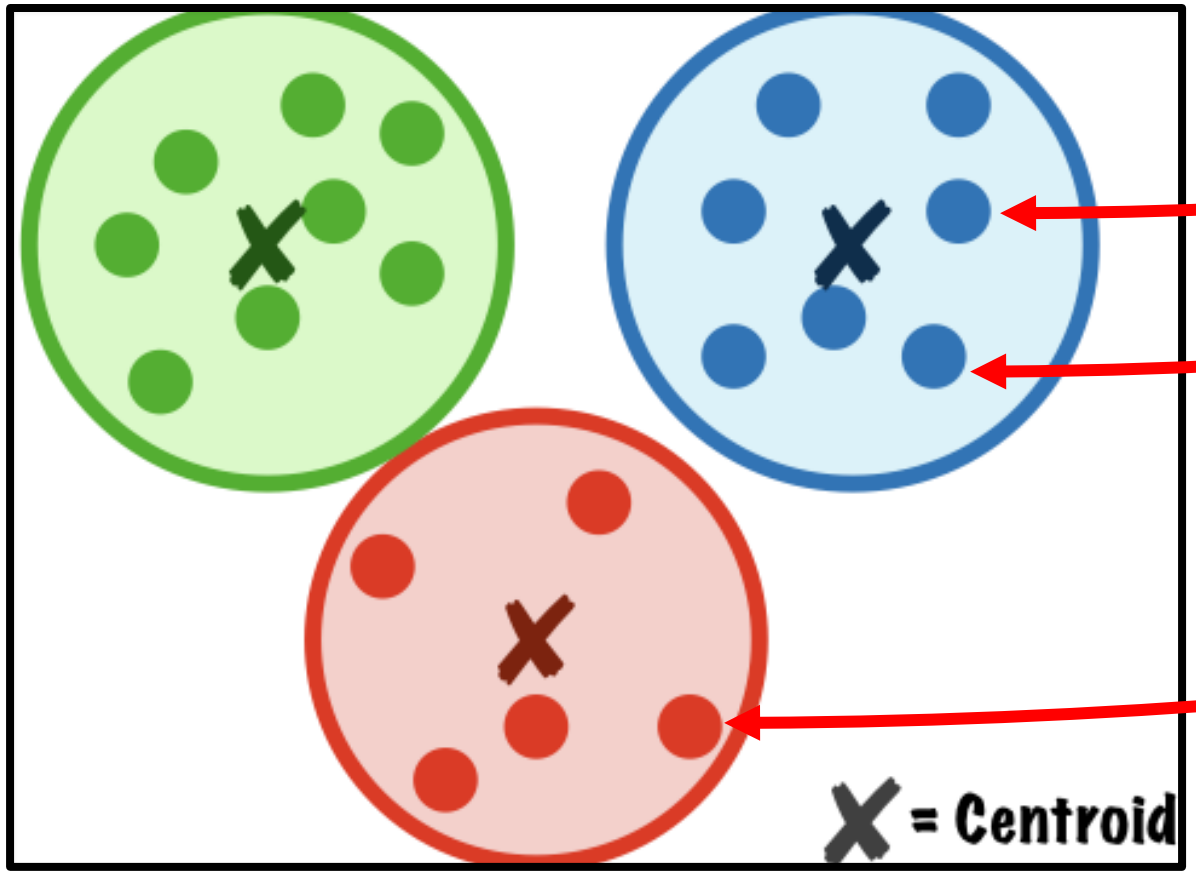
เมื่อ $t_i \in D$ ที่กำหนดให้อยู่ในคลัสเตอร์ใดคลัสเตอร์หนึ่งที่ $c_i \in C$

สำหรับทุกคลัสเตอร์ $c_i \in C$ และ $t_{ip}, t_{iq} \in c_i$ ทุกตัว เมื่อเทียบกับ $t_j \notin c_i$ จะได้ว่า

similarity (t_{ip}, t_{iq}) > similarity (t_{ip}, t_j) และ similarity (t_{iq}, t_j)

Definition of Clustering Problem (แบบแฟนซี)

Definition of Clustering Problem (แบบธรรมดา)

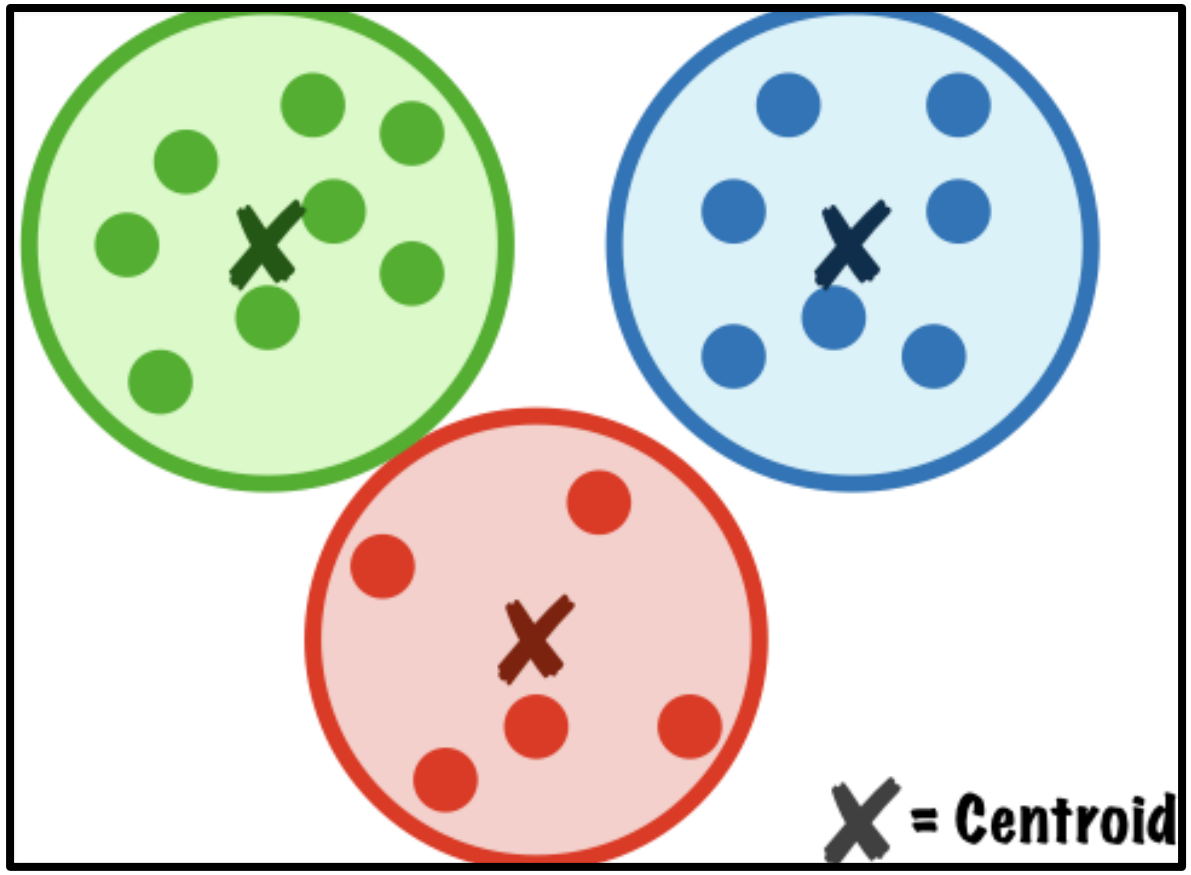


อันนี้ จะคล้ายอันนี้ แต่ไม่คล้ายอันนี้

เพราะอยู่คลัสเตอร์เดียวกัน

เพราะอยู่คนละคลัสเตอร์

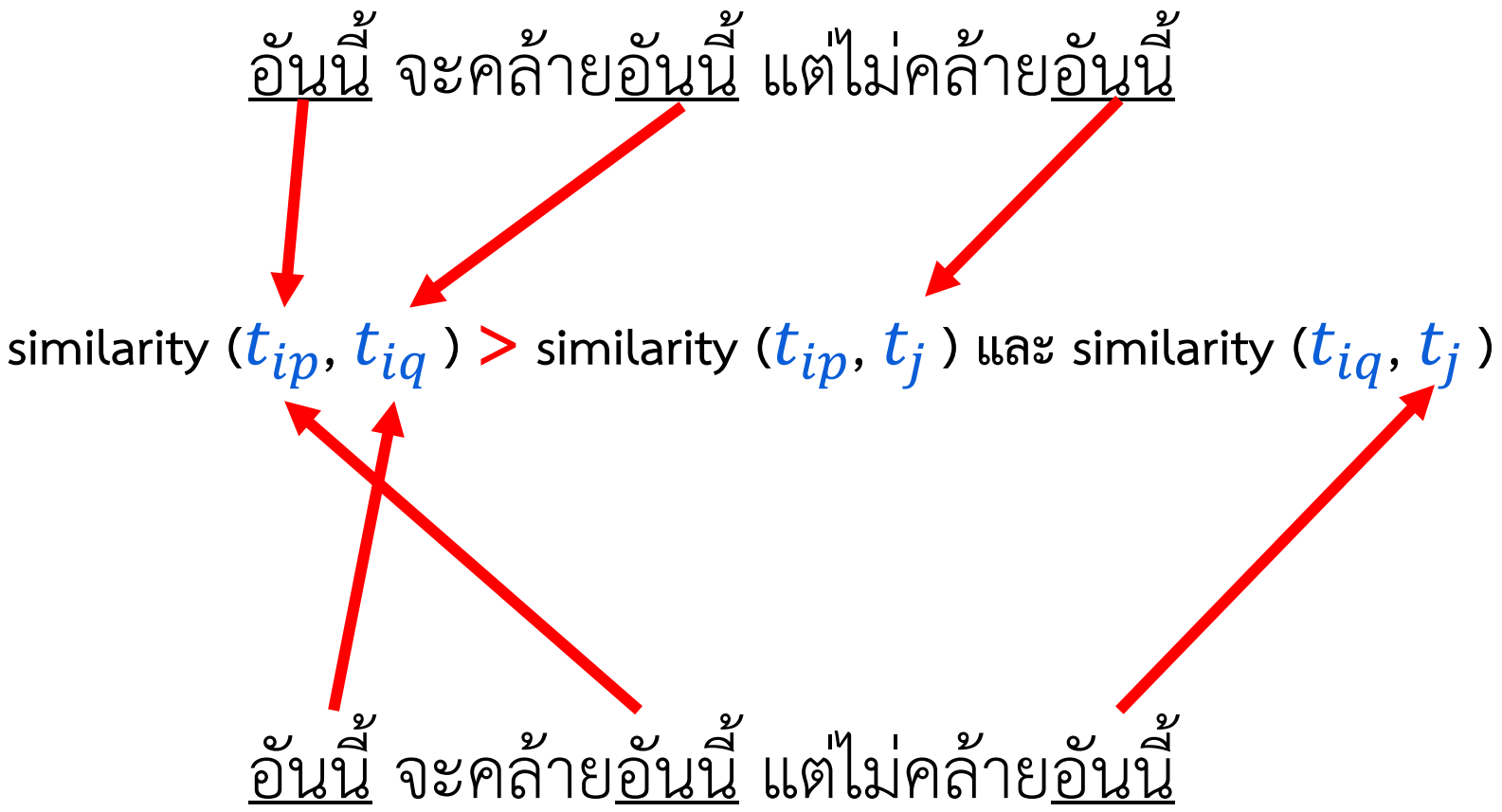
Definition of Clustering Problem (แบบธรรมดา)



ลองขีดเส้นยกตัวอย่างอื่น ๆ ที่ไม่เหมือนตัวอย่างหน้าเมื่อกี้ด้วยตัวเอง

อันนี้ จะคล้ายอันนี้ แต่ไม่คล้ายอันนี้

Definition of Clustering Problem (แบบธรรมดา)



Definition of Clustering Problem (แบบธรรมดา)

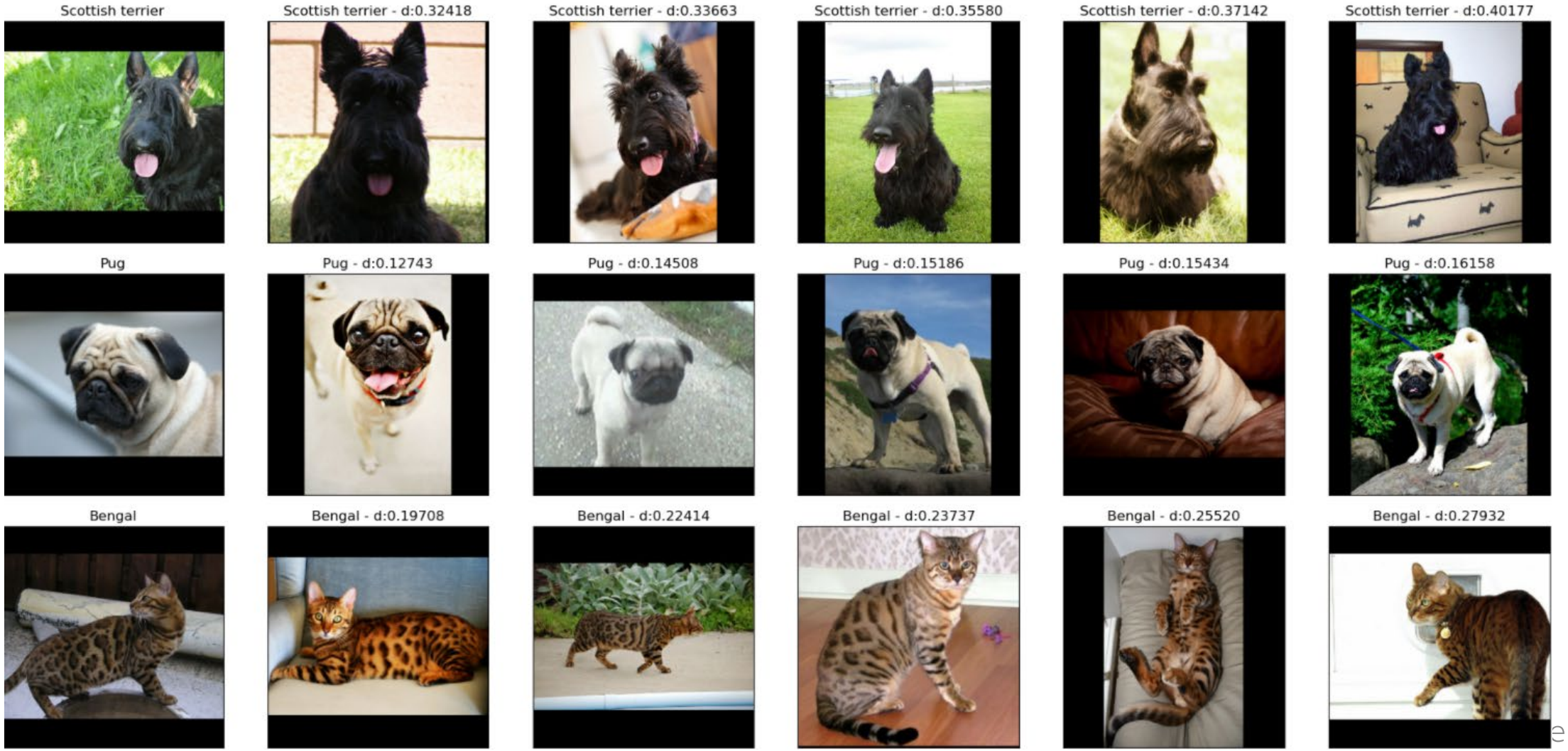
ดังนั้น ฟังก์ชัน similarity (x, y) จึงเป็นการทดสอบความคล้ายกัน (similarity) ของข้อมูลที่เป็นพารามิเตอร์ทั้งสองตัว

ในกระบวนการ Clustering ค่า similarity เป็นข้อมูลที่สำคัญ

Outline

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- Definition of Clustering Problem
- **Similarity / Dissimilarity measurement**

Similarity / Dissimilarity measurement



Similarity / Dissimilarity measurement

What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.
Webster's Dictionary

Similarity is hard to define, but…
"*We know it when we see it*"

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

Similarity / Dissimilarity measurement

- ในสายงาน Unsupervised ML นั้น การวัดค่าความเหมือน (Similarity) และความแตกต่าง (Dissimilarity) มีความสำคัญมาก
- ความหมายของทั้งสองคำนั้นตรงตัว คือ Similarity ใช้ในการวัดความเหมือนของวัตถุมากกว่า 1 ชิ้นขึ้นไป ส่วน Dissimilarity ใช้ในการวัดความแตกต่างของวัตถุมากกว่า 1 ชิ้นขึ้นไป
- เนื่องจากทั้งสองค่านี้ถูกใช้งานในลักษณะเดียวกัน เราจึงสามารถให้นิยามทั้งสองคำนี้ร่วมกันเป็นคำเดียวว่า Proximity (แปลว่า ความใกล้ชิด)

Similarity / Dissimilarity measurement

- ค่า Proximity จะมีค่ามากกว่าหรือเท่ากับ 0 เสมอ โดยที่
  - ค่า Proximity จะมีค่าเป็นมาก (อาจเป็น 1 หรือ +inf) เมื่อวัตถุทั้งสองชิ้นเหมือนกันทุกประการ หรือเป็นวัตถุชิ้นเดียวกัน (highly similar)
  - ค่า Proximity จะมีค่าเป็น 0 เมื่อวัตถุทั้งสองชิ้นแตกต่างกันโดยสิ้นเชิง (highly dissimilar)

- บ่อยครั้ง คำว่า distance มักถูกใช้เรียกแทนคำว่า dissimilarity แต่ในความจริงแล้ว distance ใช้อ้างถึง dissimilarity ที่เป็นกรณีพิเศษเท่านั้น ไม่ได้มีความหมายเหมือนกับ dissimilarity ทั้งหมด **(distance – special case of dissimilarity)**

Outline

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- Definition of Clustering Problem
- **Similarity / Dissimilarity measurement: Nominal**, *Ordinal*
  - **Single attribute**
  - Two or more attribute
  - Categorical attribute

Proximity: Single attribute

- สมมติว่ามีวัตถุอยู่จำนวนหนึ่ง แต่ละชิ้นมี Attribute เดียว (คิดง่าย ๆ ให้เป็นความยาว)

$$a_1, a_2, \ldots\ldots, a_n$$

- “Dissimilarity matrix” จะถูกสร้างขึ้นเพื่อเก็บความแตกต่างของ Attribute
- โดย Dissimilarity matrix จะเป็นเมทริกซ์ขนาด n x n

$$\begin{bmatrix} 0 & & & \\ p_{(2,1)} & 0 & & \\ p_{(3,1)} & p_{(3,2)} & 0 & \\ \vdots & \vdots & \vdots & \\ p_{(n,1)} & p_{(n,2)} & \ldots\ldots & 0 \end{bmatrix}_{n \times n}$$

- โดยที่ $p_{(i,j)}$ แทนค่า proximity ของวัตถุ 2 ชิ้น ที่มีค่า attribute เป็น a_i, a_j
- Note: Proximity measure เป็นการวัดที่ให้ผลแบบ**สมมาตร (symmetric)** กล่าวคือ $p_{(i,j)} = p_{(j,i)}$

Proximity: Single attribute

- การคำนวณ proximity (เพื่อใส่ในแต่ละตำแหน่งของ dissimilarity matrix) นั้นแตกต่างไปตามชนิดของข้อมูล (NOIR topology)
  - **Nominal attribute** -> ซ้ายขวา สี กรุ๊ปเลือด เพศ ชนิดของสิ่งของ สายพันธุ์ ราศี เป็นต้น (วันนี้)
  - **Ordinal attribute** -> เกรด คุณภาพ ขนาด (s, m, l) เป็นต้น คือข้อมูลที่สามารถระบุลำดับได้ (สัปดาห์หน้า)

Proximity: Nominal attribute; single

- จากความรู้เดิม เราทราบว่า
  - ค่า Proximity จะมีค่าเป็นมาก (อาจเป็น 1 หรือ +inf) เมื่อวัตถุทั้งสองชิ้นเหมือนกันทุกประการ หรือเป็นวัตถุชิ้นเดียวกัน (highly similar)
  - ค่า Proximity จะมีค่าเป็น 0 เมื่อวัตถุทั้งสองชิ้นแตกต่างกันโดยสิ้นเชิง (highly dissimilar)

Proximity: Nominal attribute; single

- ดังนั้น จงพิจารณาข้อมูลต่อไปนี้

Object	Gender
Ram	Male
Sita	Female
Laxman	Male

จะกล่าวได้ว่า

$$p(Ram, sita) = 0$$

$$p(Ram, Laxman) = 1$$

ในกรณีนี้ ถ้ากำหนดให้ q แทน dissimilarity ระหว่าง 2 วัตถุ i และ j ที่เป็น Single attribute จะได้ว่า

$$q_{(i,j)} = 1 - p_{(i,j)}$$

Outline

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- Definition of Clustering Problem
- **Similarity / Dissimilarity measurement : Nominal**, *Ordinal*
  - Single attribute
  - **Two or more attribute**
  - Categorical attribute

Proximity: Nominal attribute; two or more

- สมมติว่ามีจำนวน Attribute เป็น b เราจะสามารถสร้าง contingency table เพื่อสรุปความเหมือนหรือแตกต่างกันของวัตถุ x และ y ได้ดังนี้

Object x	Object y	
	1	0
1	f_{11}	f_{10}
0	f_{01}	f_{00}

เมื่อ f_{11}= จำนวนของ attribute เมื่อ x=1 and y=1.

f_{10}= จำนวนของ attribute เมื่อ x=1 and y=0.

f_{01}= จำนวนของ attribute เมื่อ x=0 and y=1.

f_{00}= จำนวนของ attribute เมื่อ x=0 and y=0.

$$f_{00} + f_{01} + f_{10} + f_{11} = b$$

Proximity: Nominal attribute; two or more

- เช่น Gender = {M, F}
  Food = {V, N}
  Caste = {H, M}
  Education = {L, I}
  Hobby = {T, C}
  Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

Hari \ Ram		0	1
	0	2	2
	1	1	1

Proximity: Nominal attribute; two or more

- เช่น Gender = {M, F}
  Food = {V, N}
  Caste = {H, M}
  Education = {L, I}
  Hobby = {T, C}
  Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

Tomi	Ram	
	0	1
0		
1		

Tomi	Hari	
	0	1
0		
1		

Proximity: Nominal attribute; two or more

เราสามารถวัดค่า proximity จากข้อมูลนี้ได้ 2 แบบ ได้แก่

(1) Symmetric binary attribute (attribute ไบนารีสมมาตร)

(2) Asymmetric binary attribute (attribute ไบนารีไม่สมมาตร)

Proximity: Nominal attribute; two or more

- ในการวัด similarity ระหว่างสองวัตถุที่มี attribute มากกว่า 1 โดยวิธีไบนารีสมมาตร เราจะใช้วิธีการวัดที่ชื่อว่า symmetric binary coefficient หรือ $\mathcal{S}$

$$\mathcal{S} = \frac{Number\ of\ matching\ attribute\ values}{Total\ number\ of\ attributes}$$

หรือ

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Proximity: Nominal attribute; two or more

- ในการวัด dissimilarity ระหว่างสองวัตถุที่มี attribute มากกว่า 1 โดยวิธีไบนารีสมมาตร เราจะใช้วิธีการวัดที่ตรงข้ามกับ symmetric binary coefficient หรือ $\mathcal{D}$

$$\mathcal{D} = \frac{Number\ of\ mismatched\ attribute\ values}{Total\ number\ of\ attributes}$$

หรือ

$$\mathcal{D} = \frac{f_{01} + f_{10}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Note that, $\mathcal{D} = 1 - \mathcal{S}$

Proximity: Nominal attribute; two or more

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

Hari \ Ram		0	1
		0	1
0		2	2
1		1	1

$$\mathcal{S}(\text{Hari, Ram}) = \frac{1+2}{1+2+1+2} = 0.5$$

$\mathcal{S}$(Hari, Tomi) =

$\mathcal{S}$(Ram, Tomi) =

Gender	= {M, F}
Food	= {V, N}
Caste	= {H, M}
Education	= {L, I}
Hobby	= {T, C}
Job	= {Y, N}

Proximity: Nominal attribute; two or more

- ในการวัด similarity ระหว่างสองวัตถุที่มี attribute มากกว่า 1 โดยวิธีไบนารีไม่สมมาตร (asymmetric) เราจะใช้วิธีการวัดที่ชื่อว่า Jaccard coefficient หรือ $\mathcal{J}$

$$\mathcal{J} = \frac{Number\ of\ matching\ presence}{Number\ of\ attributes\ not\ involved\ in\ 00\ matching}$$

หรือ

$$\mathcal{J} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Proximity: Nominal attribute; two or more

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

Hari \ Ram	0	1
0	2	2
1	1	1

$$\mathcal{J}(\text{Hari, Ram}) = \frac{1}{2+1+1} = 0.25$$

Gender	= {M, F}
Food	= {V, N}
Caste	= {H, M}
Education	= {L, I}
Hobby	= {T, C}
Job	= {Y, N}

Note: $\mathcal{J}(\text{Hari, Ram}) = \mathcal{J}(\text{Ram, Hari})$

Proximity: Nominal attribute; two or more

- จากนั้น สร้าง similarity matrix โดยใช้ Jaccard coefficient สำหรับวัตถุทั้งหมด

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

$$\mathcal{J} = \begin{array}{c} \\ H \\ R \\ T \end{array} \begin{array}{c} \begin{array}{ccc} H & R & T \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ \mathcal{J}(R,H) & 0 & 0 \\ \mathcal{J}(T,H) & \mathcal{J}(T,R) & 0 \end{bmatrix} \end{array}$$

Outline

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- Definition of Clustering Problem
- **Similarity / Dissimilarity measurement : Nominal**, *Ordinal*
  - Single attribute
  - Two or more attribute
  - **Categorical attribute**

Proximity: Nominal attribute; categorical

- ที่ผ่านมาเรากล่าวถึง attribute ที่มีลักษณะที่เป็นไปได้เพียง 2 รูปแบบต่อหนึ่ง attribute เท่านั้น
- Categorical attribute ก็เป็น attribute แบบ nominal อีกลักษณะหนึ่ง ที่หนึ่ง attribute มีลักษณะที่เป็นไปได้มากกว่า 1 รูปแบบ
  - เช่น สี {แดง, เขียว, น้ำเงิน, เหลือง}

Proximity: Nominal attribute; categorical

- ถ้ากำหนดให้ $s(x, y)$ แทน similarity ระหว่างวัตถุ x และ y แล้ว

$$s(x, y) = \frac{Number\ of\ matches}{Total\ number\ of\ attributes}$$

- ในทางกลับกัน dissimilarity คือ $d(x, y)$ โดยที่

$$d(x, y) = \frac{Number\ of\ mismatches}{Total\ number\ of\ attributes}$$

Proximity: Nominal attribute; categorical

- ย่อสั้น ๆ ได้ว่า

$$s(x,y) = \frac{m}{a} \quad \text{และ} \quad d(x,y) = \frac{a-m}{a}$$

เมื่อ a แทนจำนวน attribute ทั้งหมด

และ m แทนจำนวน attribute ที่ตรงกัน

Proximity: Nominal attribute; categorical

- ตัวอย่าง

Object	Color	Position	Distance
1	R	L	L
2	B	C	M
3	G	R	M
4	R	L	H

- ถ้าเราสนใจเพียง attribute สีเพียงอย่างเดียว จะสร้าง similarity matrix ได้ว่า

$$s = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Dissimilarity matrix, $d =$?

Proximity: Nominal attribute; categorical

- **แบบฝึกหัด:** จงสร้าง dissimilarity matrix โดยพิจารณา categorical attributes ทั้งหมด (เช่น สี, ตำแหน่ง, ระยะทาง)

Object	Color	Position	Distance
1	R	L	L
2	B	C	M
3	G	R	M
4	R	L	H

สรุป

- แนวคิดและแนวทางการประยุกต์ใช้ Clustering
- ปัญหาที่พบบ่อยของกระบวนการ Clustering ในชีวิตจริง
- Definition of Clustering Problem
- Similarity / Dissimilarity measurement : Nominal, *Ordinal*
  - Single attribute
  - Two or more attribute
  - Categorical attribute