ความเหมือน ความไม่เหมือน

Similarity, Dissimilarity, and Proximity

ค่ามันจะตรงข้ามกัน

- **Similarity measure** or **similarity function** สร้าง Function similarity เพื่อวัดว่า input 2 ตัวเหมือน / ต่าง กันยังไง
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike

    มีค่า < 1 0 = ไม่เหมือนกัน 1 = เหมือนกัน
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- **Dissimilarity** (or **distance**) measure

  ระยะห่าง

  ถ้าจุด 2 จุดห่างกันมากเท่าไหร่ยิ่งต่างมาก ห่างมากต่างมาก
  - Numerical measure of how different two data objects are
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar)

    เหมือน
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity

  ระยะห่างระหว่าง Data เหมือนกัน

Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with *l* dimensions

$$D = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1l} \\ x_{21} & x_{22} & \dots & x_{2l} \\ \boxed{?} & \boxed{?} & \boxed{?} & \boxed{?} \\ x_{n1} & x_{n2} & \dots & x_{nl} \end{pmatrix}$$

- Dissimilarity (distance) matrix
  - n data points, but registers only the distance *d*(*i, j*) (typically metric)
  - Usually symmetric, thus a triangular matrix
  - Distance functions are usually different for real, boolean, categorical, ordinal, ratio, and vector variables
  - Weights can be associated with different variables based on applications and data semantics

$$\begin{pmatrix} 0 & & \\ d(2,1) & 0 & \\ \boxed{?} & \boxed{?} & \boxed{?} \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

Standardizing Numeric Data

- Z-score: $z = \frac{x-\mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the mean absolute deviation

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

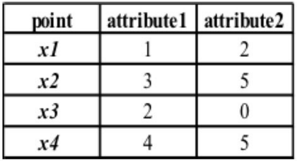
where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

  - standardized measure (*z-score*): $z_{if} = \frac{x_{if} - m_f}{s_f}$
- Using mean absolute deviation is more robust than using standard deviation

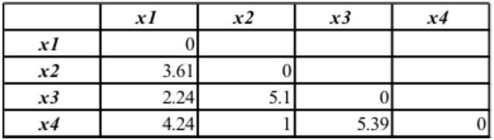
Example: Data Matrix and Dissimilarity Matrix

Data Matrix



point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (by Euclidean Distance)



	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- Minkowski distance: A popular distance measure

$$d(i,j)=\sqrt[p]{|x_{i_1}-x_{j_1}|^p+|x_{i_2}-x_{j_2}|^p+\boxed{?}+|x_{i_l}-x_{j_l}|^p}$$

where *i* = (*x*i1, *x*i2, …, *x*il) and *j* = (*x*j1, *x*j2, …, *x*jl) are two *l*-dimensional data objects, and *p* is the order (the distance so defined is also called L-*p* norm)

- Properties
  - d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positivity)
  - d(i, j) = d(j, i) (Symmetry)
  - d(i, j) □ d(i, k) + d(k, j) (Triangle Inequality)
- A distance that satisfies these properties is a metric
- Note: There are nonmetric dissimilarities, e.g., set differences

Special Cases of Minkowski Distance

- $p = 1$: (L1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \boxed{?} + |x_{il} - x_{jl}|$$

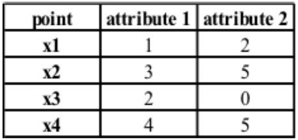
- $p = 2$: (L2 norm) Euclidean distance

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \boxed{?} + |x_{il} - x_{jl}|^2}$$

- p □ □: (Lmax norm, L□ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

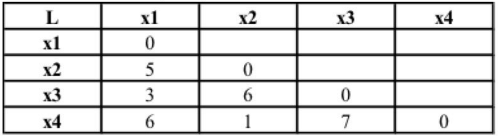
$$d(i,j) = \lim_{p \to \infty} \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p} = \max_{f=1}^{l} |x_{if} - x_{jf}|$$

Example: Minkowski Distance at Special Cases



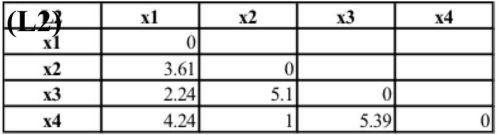
point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Manhattan (L1)



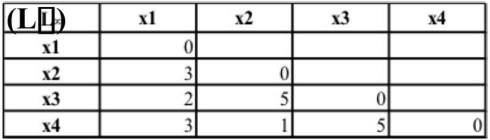
L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L2)



L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

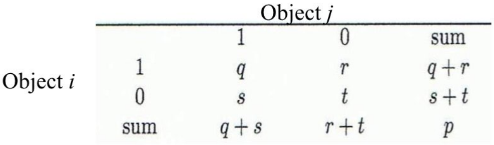
Supremum (L∞)



L∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Proximity Measure for Binary Attributes

- A contingency table for binary data



		Object *j*		
		1	0	sum
Object *i*	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables $$d(i, j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables: $$d(i, j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables): $$sim_{Jaccard}(i, j) = \frac{q}{q+r+s}$$

- Note: Jaccard coefficient is the same as "coherence" (a concept discussed in Pattern Discovery)

$$coherence(i, j) = \frac{sup(i, j)}{sup(i) + sup(j) - sup(i, j)} = \frac{q}{(q+r) + (q+s) - q}$$

Example: Dissimilarity between Asymmetric Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute (not counted in)
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- Distance: $d(i, j) = \frac{r + s}{q + r + s}$

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Mary

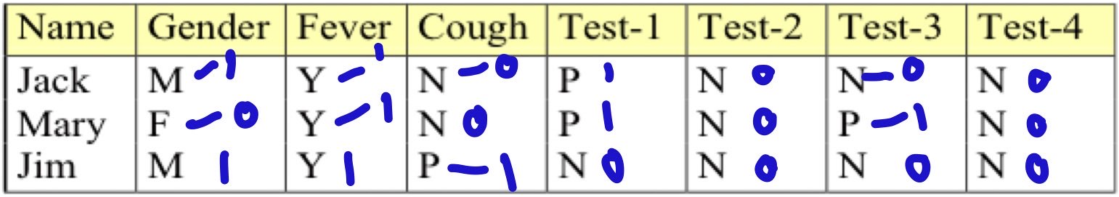
Jack	1	0	∑row
1	2	0	2
0	1	3	4
	3	3	6

Jim

Jack	1	0	∑row
1	1	1	2
0	1	3	4
∑c	2	4	6

Mary

Jim	1	0	∑row
1	1	1	2
0	2	2	4
∑c	3	3	6



Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M 1	Y 1	N 0	P 1	N 0	N 0	N 0
Mary	F 0	Y 1	N 0	P 1	N 0	P 1	N 0
Jim	M 1	Y 1	P 1	N 0	N 0	N 0	N 0

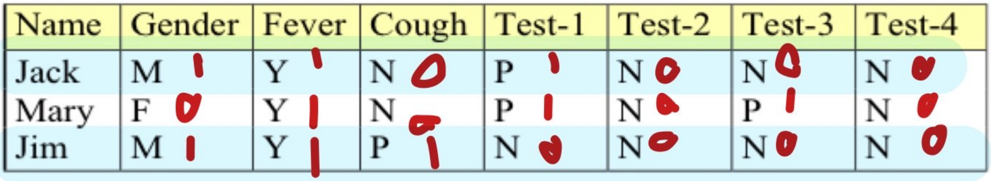
Mary

		1	0	Sum
Jack	1	2 q	1 r	3
	0	1 s	3 t	4
	Sum	3	4	7

distance . ความไม่เหมือน

symmetric $d(i,j) = \dfrac{r+s}{q+r+s+t} = \dfrac{1+1}{7} = \dfrac{2}{7}$

asymmetric $d(i,j) = \dfrac{r+s}{q+r+s} = \dfrac{1+0}{2+0+1} = \dfrac{2}{3}$



Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M 1	Y 1	N 0	P 1	N 0	N 0	N 0
Mary	F 0	Y 1	N 0	P 1	N 0	P 1	N 0
Jim	M 1	Y 1	P 1	N 0	N 0	N 0	N 0

	1	0	sum
1	1 q	1 r	2
0	1 s	3 t	4
sum	2	4	6

$$d(i,j) = \frac{r+s}{q+r+s+t} = \frac{2}{4}$$

asymmetric → $\frac{2}{3}$

Proximity Measure for Categorical Attributes

ชื่อ

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.
- <u>Method 1</u>: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i,j) = \frac{p-m}{p} \qquad \frac{\text{จำนวนตัวที่ไม่เหมือน}}{\text{จำนวนทั้งหมด}}$$

- <u>Method 2</u>: Use a large number of binary attributes
  - Creating a new binary attribute for each of the *M* nominal states

สี	อาชีพ
r	นศ
r	อจ
g	นศ

↓ r, g, b

↓ ว, นศ, อจ, Grab

	สีR	สีG	สีB	ว่?	นศ	อจ	Grab
1	1	0	0	0	1	0	0
	1	0	0	0	0	1	0
3	0	1	C	0	1	0	0

$$\text{binary} = \frac{2}{7} = \frac{\text{แตกต่าง}}{\text{ทั้งหมด}}$$