ความเหมือน ความไม่เหมือน

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function** → 2 จุดเหมือนกันอย่างไร → ค่าออกมาเป็น [0, 1] ไม่เหมือน เหมือน
  - A real-valued function that quantifies the similarity between two objects
  - Measure how two data objects are alike: The higher value, the more alike
  - Often falls in the range [0,1]: 0: no similarity; 1: completely similar
- **Dissimilarity** (or **distance**) measure ⇒ ความไม่เหมือน / ระยะห่าง
  - 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) → เหมือนกันเป็น 0
  - (เหมือน, ต่าง) →ห่างเรื่อยๆ
  - Range [0, 1] or [0, ∞) , depending on the definition
- **Proximity** usually refers to either similarity or dissimilarity → ความต่าง*

‹#›

Data Matrix and Dissimilarity Matrix

- Data matrix
  - A data matrix of n data points with l dimensions
- Dissimilarity (distance) matrix → ใช้คำนวณ ก่อนประเมินผล
  - 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

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

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

‹#›

Standardizing Numeric Data ควรปรับ Scale ให้เท่ากันก่อน

- Z-score: $z = \dfrac{x - \mu}{\sigma}$
  - 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

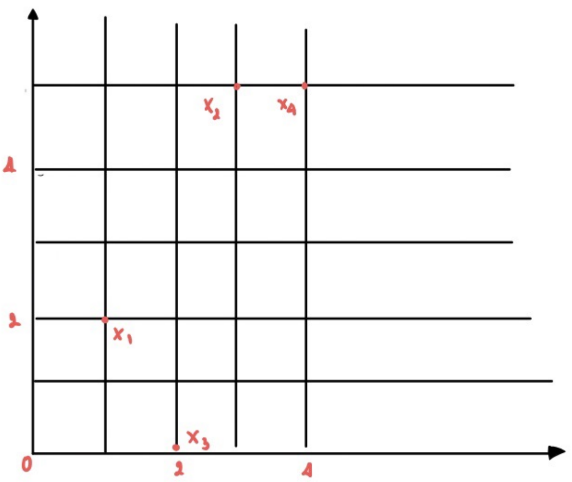
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

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

  - standardized measure (*z-score*): $z_{if} = \dfrac{x_{if} - m_f}{s_f}$
- 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

คู่เดียวกัน

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

ค่าเดียวกัน

Special Cases of Minkowski Distance → 3 แบบ

- $p = 1$: (L1 norm) Manhattan (or city block) distance → ตามแนวแกน ไม่วัดแนวเฉียง
  - 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 → $c^2 = a^2 + b^2$ → วัดความห่างของจุด

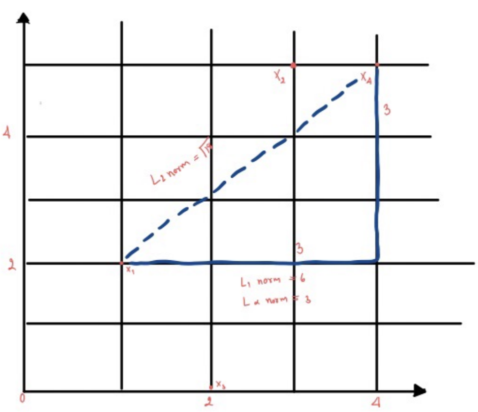
$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \boxed{?} + |x_{il} - x_{jl}|^2}$$

- $p \to \infty$: (Lmax norm, L∞ norm) "supremum" distance → ดูแต่ละแกน → ดูแกนที่มากที่สุด
  - 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