Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Similarity, Dissimilarity, and Proximity

- **Similarity measure** or **similarity function** จะสร้างฟังก์ชันขึ้นมา เพื่อดูว่าเหมือนกันอย่างไร
  - 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 (ไม่เหมือนกันเลย); 1: completely similar (เหมือนกัน)
- **Dissimilarity** (or **distance**) measure ไม่เหมือน
  - Numerical measure of how different two data objects are ระยะห่างระหว่างจุดสองจุด ห่างกันมาก แปลว่าต่างกันมาก
  - In some sense, the inverse of similarity: The lower, the more alike
  - Minimum dissimilarity is often 0 (i.e., completely similar) [0,1] ; เหมือนกัน
  - Range [0, 1] or [0, ∞) , depending on the definition [0,∞] ; ห่างกัน
- **Proximity** usually refers to either similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

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

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

- 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

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

วัดความเหมือน หรือต่างกันของ Data

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

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

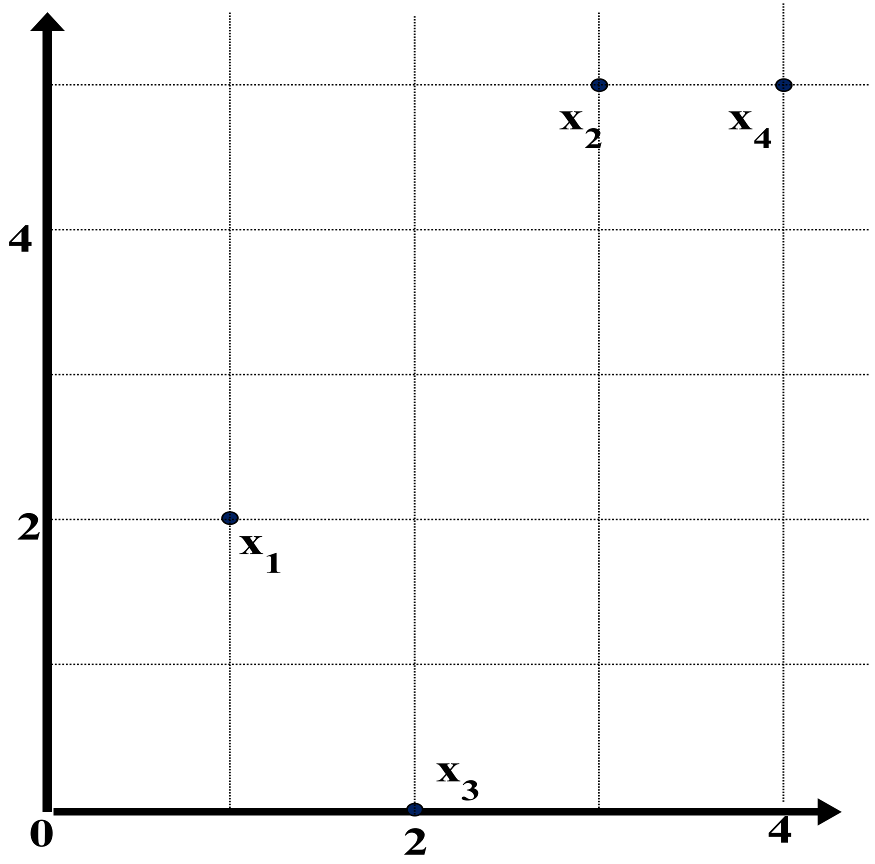
where

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

ถ้าค่าใน column ไม่เท่ากันควรปรับ scale
ให้เท่ากัน ไม่งั้นจะเกิดการไบแอสกับ column ใดอันนึง

  - 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

$$d(i, j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{il})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jl})$ 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: (L_1 norm) Manhattan (or city block) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

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

- p = 2: (L_2 norm) Euclidean distance

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

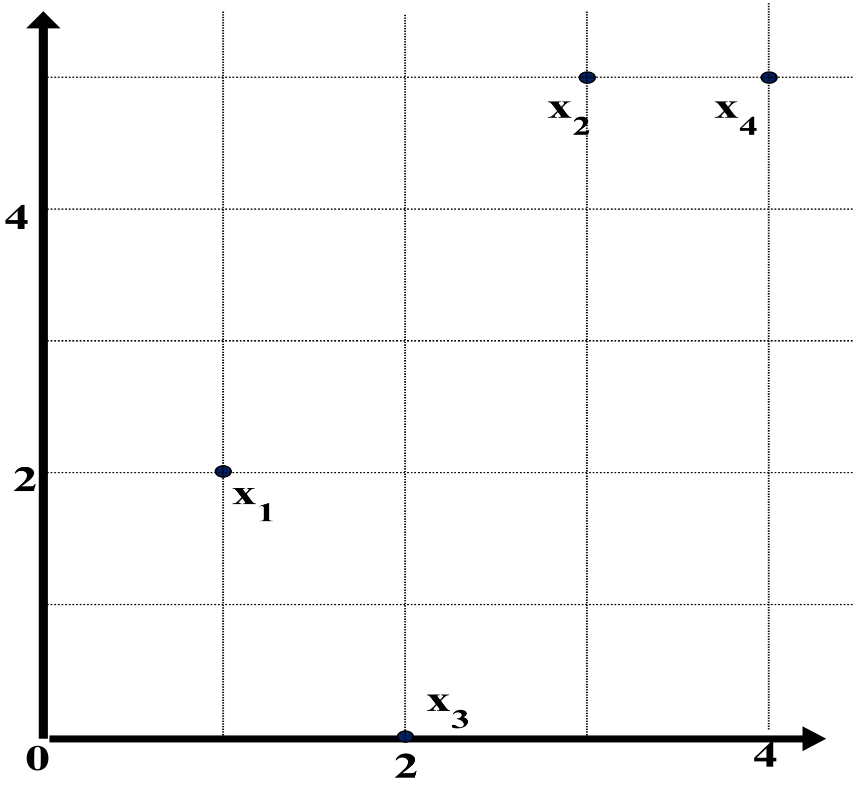
- $p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance
  - The maximum difference between any component (attribute) of the vectors

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

สูตรนี้จะนำไปคำนวณ Data

Example: Minkowski Distance at Special Cases

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



Manhattan (L_1)

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

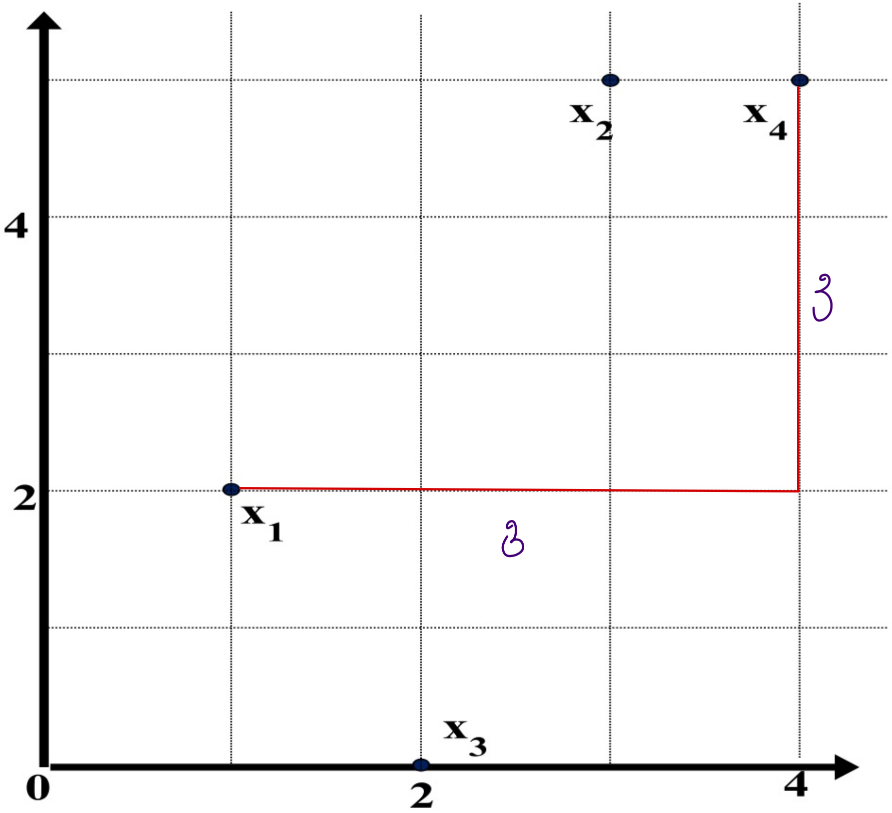
Euclidean (L_2)

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

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

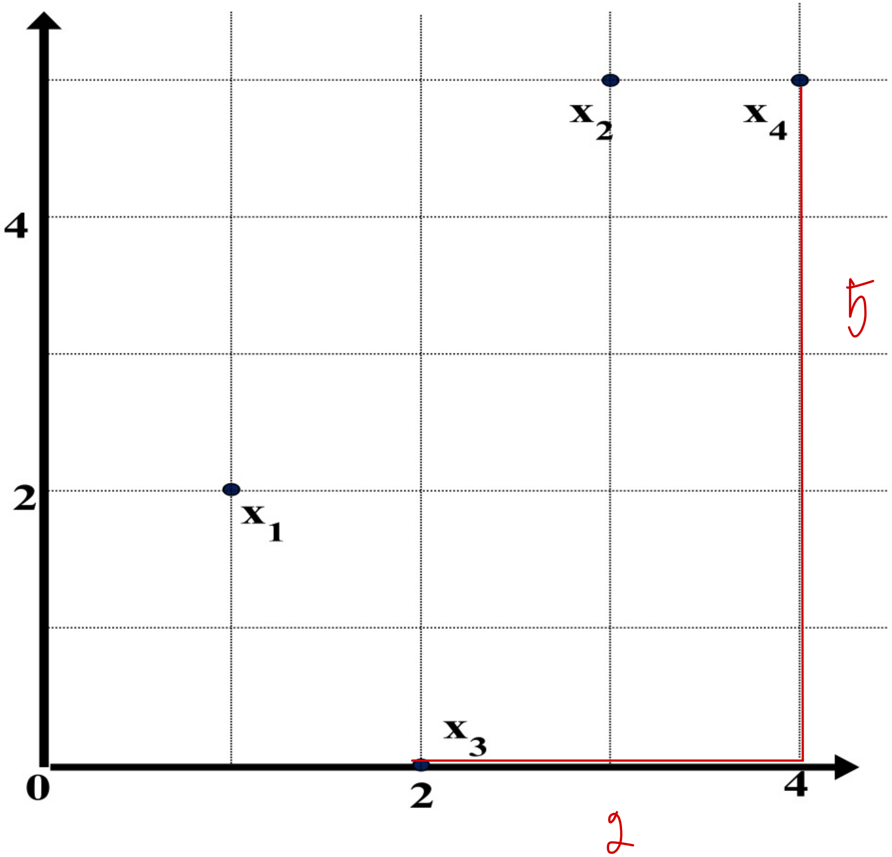


L_1 norm $= 6 = 3+3$

L_2 norm $= \sqrt{18} = \sqrt{3^2+3^2}$

$L_\infty = 3 =$ เส้นที่มีค่ามากที่สุด

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

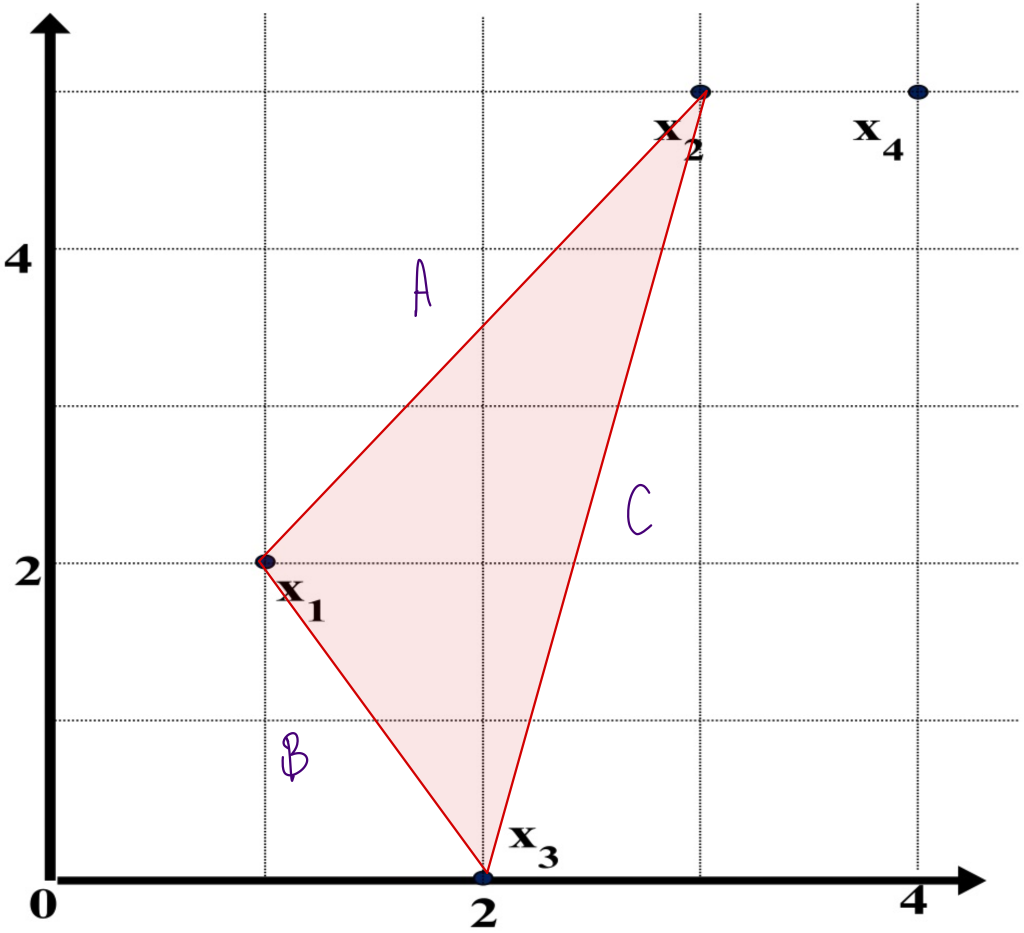


L1 norm = 7 = 5+2

L2 norm = $\sqrt{29}$ = $\sqrt{5^2 + 2^2}$

L_∞ = 5 = max

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



ระยะทางระหว่างเส้น A + เส้น B

จะมากกว่า ระยะทางเส้น C เสมอ