สรุป

Chapter 2

Measuring Data Similarity and Disimilarity

- เป็นการวัด Data เพื่อเปรียบเทียบข้อมูลที่เหมือนกัน ต่างกัน

1. Similarity measure or similarity function
   - เป็นการวัดค่าของข้อมูลที่เหมือนกัน / คล้ายกัน
   - โดยอยู่ในช่วง 0, 1 ; 0 = ไม่เหมือนกัน
     1 = เหมือนกัน

2. Dissimilarity (or distance) measure
   - เป็นการวัดความแตกต่างระหว่างข้อมูล เช่น คุณ 2 คุณ ต่างกัน ห่างกัน มากเท่าไหร่ โดยค่าจะอยู่ระหว่าง $[0, 1]$ หรือ $[0, \infty]$ ขึ้นอยู่กับค่าของข้อมูล

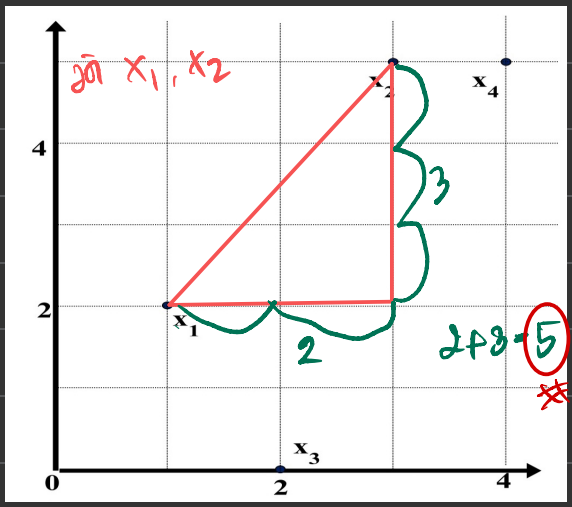
3. Poximity ความใกล้เคียงของคุณของข้อมูล

การวัดระยะห่าง เมื่อรู้จัก Distance matrix

1 L_1 norm เป็นการวัดระยะแบบชายตึก

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

Ex



2. L_2 norm เป็นการวัดโดยใช้สูตร ฟีทาโกรัส

$A^2 + B^2 = C^2$

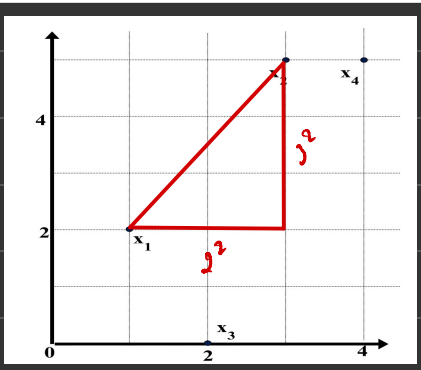
2. L_2 norm เป็นการวัดโดยใช้สูตร พีทาโกรัส

$A^2 + B^2 = C^2$

- $p = 2$: (L_2 norm) Euclidean distance

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

Ex จงหา x_1, x_2



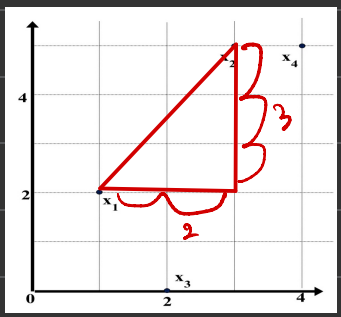
$3^2 + 3^2 = C^2 = 4 + 9$

$13 = C^2$ $\therefore C = \sqrt{13}$

3. L_∞ หรือ L_{max} norm หรือค่ามากที่สุด

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

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



เลือก จำนวนที่มากที่สุด คือ 3

Proximity for Binary

ระยะของตัวแปร Binary มี 2 แบบ คือ

- Symmetric
- asymmetric

- A contingency table for binary data

Object *i* \ Object *j*	1	0	sum
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: $d(i,j) = \frac{r+s}{q+r+s}$

ระยะของ Categorical มี 2 วิธี

1 simple matching

- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i,j) = \frac{p-m}{p}$$

m = จำนวนตัวแปรที่ตรงกัน เช่น เหมือนกัน

p = จำนวนตัวแปรทั้งหมด

2. ใช้ตัวแปรที่เป็น binary

Ordinal

ตัวแปรที่มีค่าเรียงลำดับ เช่น 0, 1

วิธีแก้ปัญหา

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

$r \in (1, \ldots, M_f)$

M_f คือ จำนวนสถานะที่เป็นไปได้