"Measuring Data Similarity and Dissimilarity"

ศิรกานต์ ร่มเรือ

6450200539

Similarity, Dissimilarity, and Proximity

- Similarity measure or similarity function

  - Similarity คือความเหมือน, Similarity Measure จะเป็นการวัดความเหมือน โดยมีค่า 0-1 เพื่อใช้เปรียบเทียบส่วนต่างของไอเทม

  ถ้าเหมือนมากค่าจะเข้าใกล้ 1 และ ดูว่า function ไหนที่จะช่วยวัดได้ดีที่สุด เช่น จุดที่ไม่เหมือนกันมากแค่ไหน

- Dissimilarity (or distance) ระยะทาง
  - การวัดว่าข้อมูลต่างกันมากแค่ไหนที่ต้องการ ซึ่งใช้ ค่า distance และ โดยถ้าเหมือนกัน ค่าจะเข้าใกล้ 0 Similarity เป็น inverse กับ distance

- Proximity

  - Proximity คือความใกล้ระหว่างจุดข้อมูล

Data Matrix and Dissimilarity Matrix

Data Matrix (ฟีเจอร์ที่ 1, ฟีเจอร์ที่ 2 — columns; data ตัวที่ 1, data ตัวที่ 2 — rows)

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

- วัดระยะห่างแล้วด้วย Matrix หรือ distance matrix

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

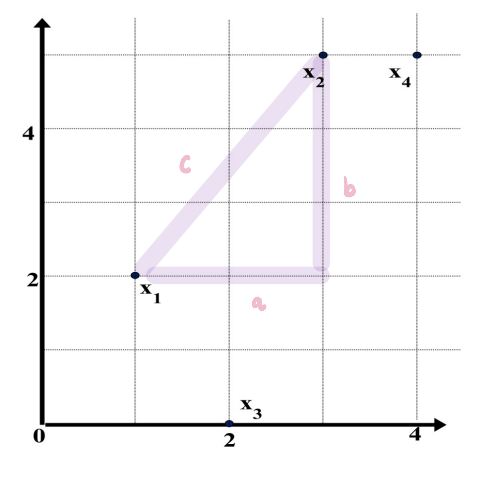
ถ้าข้อมูลมี 4 จุด distance matrix จะเท่ากับ 4x4

Standardizing Numeric Data

ต้องจัดการข้อมูลให้อยู่ในรูปแบบมาตรฐานเดียวกัน $z \sim N(0,1)$

$$z = \frac{x - \mu}{\sigma}$$

Example: Data Matrix and Dissimilarity



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

จากรูปสามเหลี่ยมมุมฉาก $c^2 = a^2 + b^2$ เราจะหาระยะห่างระหว่าง x_1 กับ x_2 เท่ากับ $c^2 = 2^2 + 3^2$ ∴ $c = 3.61$

Minkowski distance

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

กรณีที่ 1 (L_1 norm) Manhattan (or city block) distance

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

(ใช้ absolute เพื่อให้ระยะห่างที่คำนวณไม่ติดลบ)

กรณีที่ 2 (L_2 norm) Euclidean distance

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

(ยกกำลังเพื่อให้ค่าเลขไม่ติดลบ)

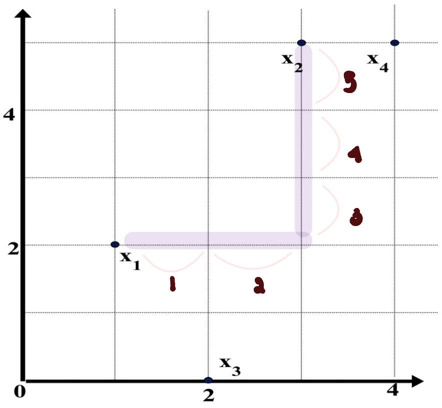
$p \to \infty$: (L_{max} norm, L_∞ norm) "supremum" distance

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

Example: Minkowski Distance at Special Cases

L_1

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

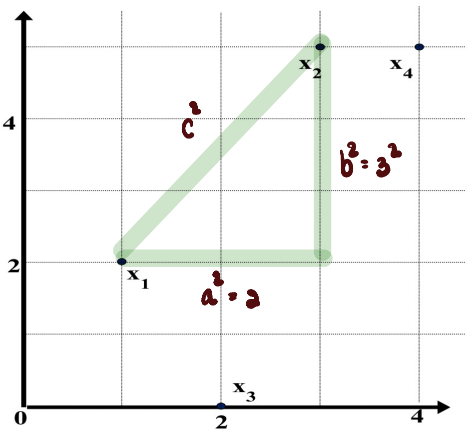
นับจำนวนช่องทั้งหมด เช่น ระยะห่างระหว่าง x_1 กับ x_2 ห่างกัน 5 จุด

โดยจุด x_1 ถึงมุม = 2 จุด และ จากมุมถึง x_2 = 3 จุด

ดังนั้น x_1 และ x_2 ห่างกัน 2+3 = 5 จุด

L_2

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



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

หาโดยใช้พีทาโกรัส $c^2 = a^2 + b^2$

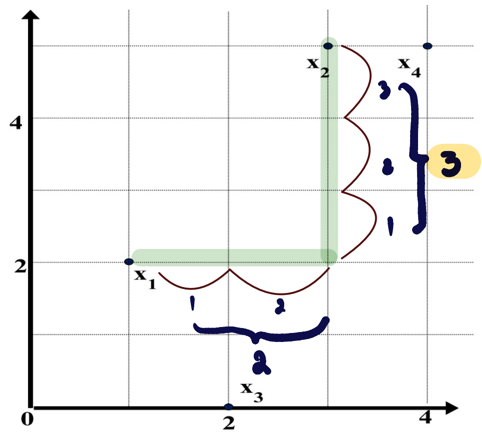
$$c^2 = 2^2 + 3^2$$

$$c = \sqrt{4 + 9}$$

ดังนั้น $c = 3.61$

L_∞

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



Supremum (L_∞)

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

จะหาค่า Max ของระยะทาง เช่น ระยะทางระหว่าง x_1 กับ x_2 โดย x_1 ถึงจุดห่าง 2 จุด และจากจุดถึง x_2 ห่างกัน 3 จุด ดังนั้น L_∞ จึง = 3 เพราะเอาค่า Max

Proximity Measure for Binary Attributes

symmetric binary (ที่ความน่าจะเป็นเท่ากัน)

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

asymmetric binary (ความน่าจะเป็นไม่เท่ากัน)

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

Object j (columns), Object i (rows)

	1	0	sum
1	q	r	q+r
0	s	t	s+t
sum	q+s	r+t	p

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

Jack

โดยค่า Y และ P=1 และ N=0

Mary / Jack

	1	0	$\sum_{row}$
1	2	0	2
0	1	3	4
$\sum_{col}$	3	3	6

เช่น มีตัวแปรที่เท่ากัน 2 ตัวคือ Y,Y ในช่อง Fever และ P,P ในช่อง Test1
จึงใส่ในช่อง (1,1) = 2

คำนวณหา $d(i,j)$ ในตัวแปร asymmetric binary

สูตร $d(\text{jack}, \text{Mary}) = \frac{r+s}{q+r+s}$

$= \frac{0+1}{2+0+1}$

$= 0.33$

Proximity Measure for Categorical Attributes

$d(i,j) = \frac{p-m}{p}$; p: total = 3 (มีแอททริบิวต์ 3 ตัวคือ จังหวัด สีที่ชอบ อาชีพ)
m: matches

	จังหวัด	สีที่ชอบ	อาชีพ
d_1	กทม	แดง	ตำรวจ
d_2	ขอนแก่น	เขียว	Programmer
d_3	กทม	น้ำเงิน	Programmer
d_4	เชียงใหม่	ส้ม	สัปเหร่อ

distance matrix = 4x4

	d_1	d_2	d_3	d_4
d_1	0			
d_2	1	0		
d_3	0.667		0	
d_4				0

d_1 กับ d_2 ไม่มีค่าที่ซ้ำกันเลย m=0

$\therefore d(d_1,d_2) = \frac{3-0}{3}$

d_1 กับ d_3 มีซ้ำกัน 1 ตัวคือ กทม m=1

$\therefore d(d_1,d_3) = \frac{3-1}{3} = \frac{2}{3} = 0.667$

dummy

จังหวัด_ตาก	จังหวัด_ขอนแก่น	จังหวัด_เชียงใหม่
1	0	0
0	1	0
1	0	0
0	0	1

Ordinal Variables

เช่น เรียงว่าห่างกันเท่าไหร่

$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \quad ; \; r_{if} \in \{1, 2, \ldots, M\}$$

Example $r_{if} \in \{1, \ldots, 4\}$

freshman : 1   Sophomore : 2   junior : 3   senior : 4

$z = \frac{1-1}{4-1} = 0$ $z = \frac{2-1}{4-1} = \frac{1}{3}$ $z = \frac{3-1}{4-1} = \frac{2}{3}$ $z = \frac{4-1}{4-1} = 1$

ถ้าต้องหาระยะห่างให้ทั้ง 2 จุด ลบกัน

$d(\text{junior}, \text{senior}) = 1 - \frac{2}{3} = \frac{1}{3}$