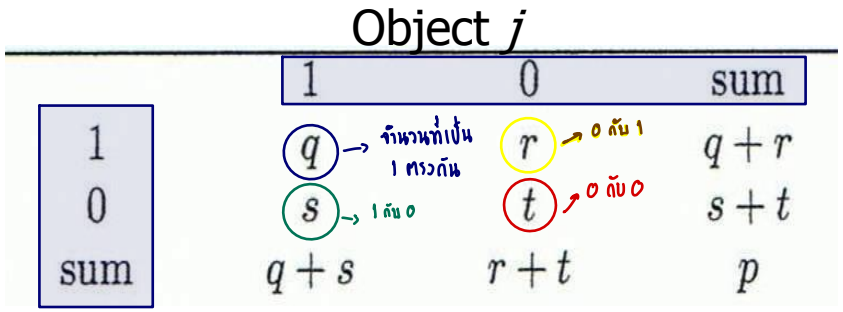
Proximity Measure for Binary Attributes

- A contingency table for binary data



- 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

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

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

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

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

Jack

	1	0	Σ_{row}
1	2	0	2
0	1	3	4
Σ_{col}	3	3	6

Mary

Jack

	1	0	Σ_{row}
1	1	1	2
0	1	3	4
Σ_{col}	2	4	6

Jim

Mary

Jim

	1	0	Σ_{row}
1	1	1	2
0	2	2	4
Σ_{col}	3	3	6

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M → 1	Y → 1	N → 0	P → Positive 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 → Negative 0	N → 0	N → 0	N → 0

mary

(Symmetric)

Jack \	1	0	sum
1	2 q	1 r	3
0	1 s	3 t	4
sum	3	4	

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

$$= \frac{1+1}{2+1+1+3} = \frac{2}{7}$$

ใช้สูตร, ทำตาราง

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M → 1	Y → 1	N → 0	P → Positive 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 → Negative 0	N → 0	N → 0	N → 0

Jack กับ Jim

ห่างกัน $\frac{2}{7}$ จากทั้งหมด

Symmetric หมายถึงไม่ใช้สูตร

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes
  - Example: Color (red, yellow, blue, green), profession, etc.

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

จน.ที่ไม่แม็ตช์ / จน.ทั้งหมด

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

m → ตัวที่แม็ตช์กัน

p − m → ต้องการหาตัวที่ไม่แม็ตช์

p → Attributes

สร้าง dummy หรือ OneHotEncoder

สี	อาชีพ
r	พ.
r	ว.
g	พ.

สีR	สีG	สีB	ว	พ.	ว.	Grab
1	0	0	0	1	0	0
1	0	0	0	0	1	0
0	1	0	0	1	0	0

r, g, b

ว, พ., ว., Grab

2/7

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

Ordinal Variables

หัวแปรอันดับ

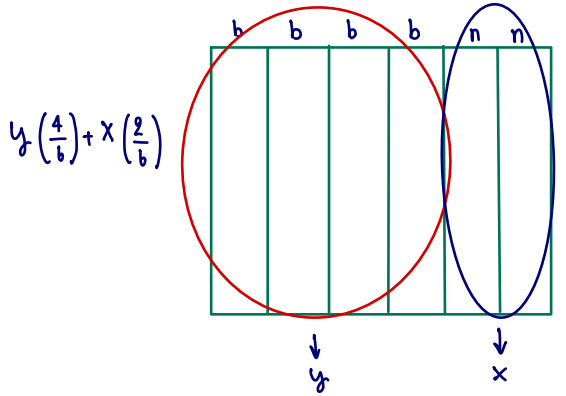
- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior)
  1 2 3 4
- Can be treated like interval-scaled
  - Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, \ldots, M_f\}$
    คือลำดับที่เท่าไหร่ของแต่ละชื่อแล้วจากน้อยมามาก
    → คือตัวสุดท้ายหลังเรียง ตย. นี้คือ 4.
  - Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by
    สูตรการหาระยะห่างระหว่างจุด 2 จุด
    หาระยะห่างด้วยการทดแต่ $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

    freshman
    $\frac{1-1}{4-1} = \frac{0}{3}$
    $= 0$
  - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
    - Then distance: d(freshman, senior) = 1, d(junior, senior) = 1/3 → หาโดยการใส่ absolute (1 — 1)
  - Compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A dataset may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

$$d(i,j) = \frac{\sum_{f=1}^{p} w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} w_{ij}^{(f)}}$$



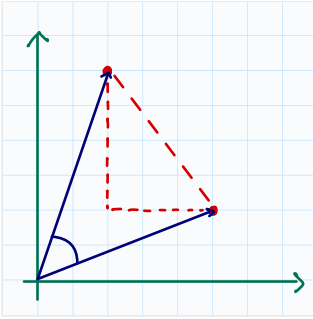
- If f is numeric: Use the normalized distance
- If f is binary or nominal: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; or $d_{ij}^{(f)} = 1$ otherwise
- If f is ordinal
  - Compute ranks z_{if} (where $z_{if} = \frac{r_{if} - 1}{M_f - 1}$)
  - Treat z_{if} as interval-scaled

Cosine Similarity of Two Vectors

นับคำศัพท์จากบทความ สเกลจะเท่าๆกัน , ไม่สามารถ normal ได้ถ้ามี document ที่เป็นนิยายเพิ่มเข้ามาเพราะคำในนิยายมันเยอะมาก, มันทำให้สเกลเท่ากันไม่ได้

- A **document** can be represented by a bag of terms or a long vector, with each attribute recording the *frequency* of a particular term (such as word, keyword, or phrase) in the document

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0



แปลงจุดเป็นเวกเตอร์ แล้วดูระยะห่างระหว่างมุม

- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

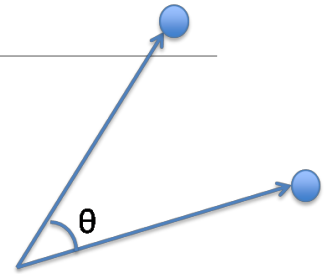
$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, $||d||$: the length of vector d

Example: Calculating Cosine Similarity

- Calculating Cosine Similarity: $cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$

$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



where • indicates vector dot product, ||*d*||: the length of vector *d*

- Ex: Find the **similarity** between documents 1 and 2.

d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

  - First, calculate vector dot product

$d_1 \bullet d_2$ = 5 X 3 + 0 X 0 + 3 X 2 + 0 X 0 + 2 X 1 + 0 X 1 + 0 X 1 + 2 X 1 + 0 X 0 + 0 X 1 = 25

  - Then, calculate ||d_1|| and ||d_2||

$$\| d_1 \| = \sqrt{5 \times 5 + 0 \times 0 + 3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0} = 6.481$$

$$\| d_2 \| = \sqrt{3 \times 3 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1} = 4.12$$

  - Calculate cosine similarity: cos(d_1, d_2) = 25/ (6.481 X 4.12) = 0.94