Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object *j*		
		1	0	sum
Object *i*	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- Distance measure for symmetric binary variables

ตอบถูก ชาย=ชาย

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

0 ตรงกับ 0
ไม่ใช่=ไม่ใช่

- Distance measure for asymmetric binary variables:

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

ชาย=หญิงวะ?

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

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

- Note: Jaccard coefficient is the same as "coherence" (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 1	Y 1	N 0	P 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 0	N 0	N 0	N 0

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

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

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

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

		Mary		
		1	0	∑row
Jack	1	2	~~0~~ 1	2
	0	1	3	4
		3	~~3~~ 4	~~6~~ 7

		Jim		
		1	0	∑row
Jack	1	~~1~~ 2	1	~~2~~ 3
	0	1	3	4
	∑c	~~2~~ 3	4	~~6~~ 7

		Mary		
		1	0	∑row
Jim	1	1	1	2
	0	2	2	4
	∑c	3	3	6

	1	0	SUM
1	q	r	
0	s	t	
SUM			

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes เป็นหมวดหมู่
  - Example: Color (red, yellow, blue, green), profession, etc.
- <u>Method 1</u>: Simple matching
  - m: # of matches, p: total # of variables

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

$$\frac{\text{ทั้งหมด - เหมือน}}{\text{ทั้งหมด}}$$

จำนวนตัวที่ไม่เหมือน

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

Ordinal Variables

- An ordinal variable can be discrete or continuous เรียงลำดับ
- 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: $r_{if} \in \{1, \ldots, M_f\}$ ลำดับที่เท่าไหร่
  - Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by $$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$ 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
  - Compute the dissimilarity using methods for interval-scaled variables $|1-0| = \left|\frac{2}{3} - \frac{3}{3}\right| = \frac{1}{3}$