Proximity Measure for Binary Attributes

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

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

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

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

Mary

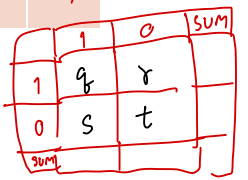
Jack	1	0	∑row
1	2	1	2
0	1	3	4
∑c	3	4	6

Jim

Jack	1	0	∑row
1	2	1	3
0	1	3	4
∑c	3	4	7

Mary

Jim	1	0	∑row
1	1	1	2
0	2	2	4
∑c	3	3	6



Proximity Measure for Categorical Attributes

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

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

ทั้งหมด - เหมือน / ทั้งหมด

ตำแหน่งตัวที่ไม่เหมือน

- Method 2: 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 $|1-0| = \left|\frac{2}{3} - \frac{3}{3}\right| = \frac{1}{3}$