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

ตัวหาร ทั้งหมด

- 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)

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

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

Mary

Jack	1	0	sum
1	2 a	1 r	3
0	1 s	3 t	4
sum	3	4	7

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

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

Jack and Jim

$= \frac{2}{7}$

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

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33 \quad \frac{1}{3}$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67 \quad \frac{2}{3}$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75 \quad \frac{3}{4}$$

Mary

Jack	1	0	Σ_{row}
1	2 q	0 r	2
0	1 s	3 t	4
Σ_{col}	3	3	6

Jim

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

Mary

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

Proximity Measure for Categorical Attributes

ชื่อ, อาชีพ

- 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

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

จำนวนตัวที่ไม่เหมือน

หารจำนวนทั้งหมด

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

สี	อาชีพ
red	น.ศ.
red	อาจารย์
green	น.ศ.
r, g, b	ทหาร, น.ศ., อาจารย์, แพทย์

	สีR	สีG	สีB	ทหาร	นศ.	อจ	แพทย์
จุดที่ 1	1	0	0	0	1	0	0
= 2/7	1	0	0	0	0	1	0
จุดที่ 3	0	1	0	0	1	0	0

Ordinal Variables

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

ลำดับที่เท่าไหร่

ระยะห่างระหว่างจุดถึงจุด

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

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

    - 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

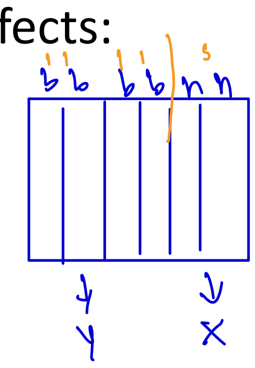
$(1-0)$ $\left(\frac{2}{3}-\frac{3}{3}\right) = \frac{1}{3}$

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:

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

$y\frac{(4)}{6} + x\frac{(12)}{(6)}$



  - 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

- 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

- 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