นางสาวอุมาพร คำภิชัย รหัสนักศึกษา 623020547-0
คณะวิทยาศาสตร์ สาขาสถิติ หลักสูตรสารสนเทศสถิติ ชั้นปีที่ 3

สรุป Chapter 2 (ต่อ)

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 (a concept discussed in Pattern Discovery)

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

** กรณี Attributes ไม่ได้เป็นตัวเลข แต่เป็น Binary

Contingency ตารางนี้ค่อยดูว่า 0 กับ 1 มันเป็นยังไง

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

Jim

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

Mary

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

- ตัวอย่าง มีจุดอยู่ 3 จุดแนวนอน Jack , Mary และ Jim
- มี 8 Attributes
- M คือ 1 F คือ 0 N เป็น 0 P เป็น 1
- Y เป็น 1 N เป็น 0

- A contingency table for binary data

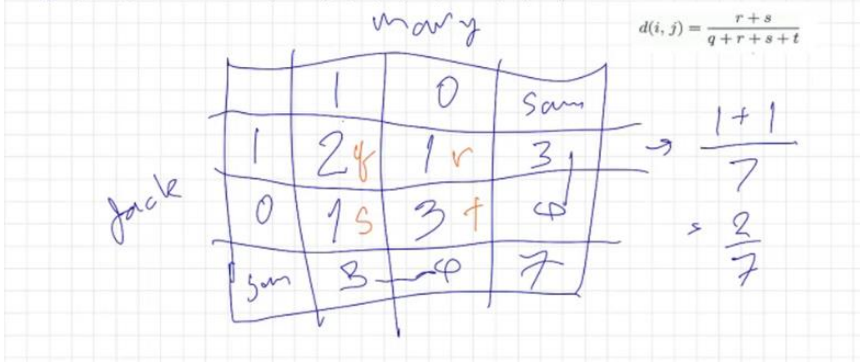
Object i \ Object j	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- 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 (a concept discussed in Pattern Discove

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

- สูตรคำนวณ
- ตัว t คือ 0 ตรงกับ 0

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



- ตารางคำนวณตัวอย่างอาจารย์ทำให้ดู เป็น Similarity
- ระยะห่างสูงสุดเท่ากับ 1

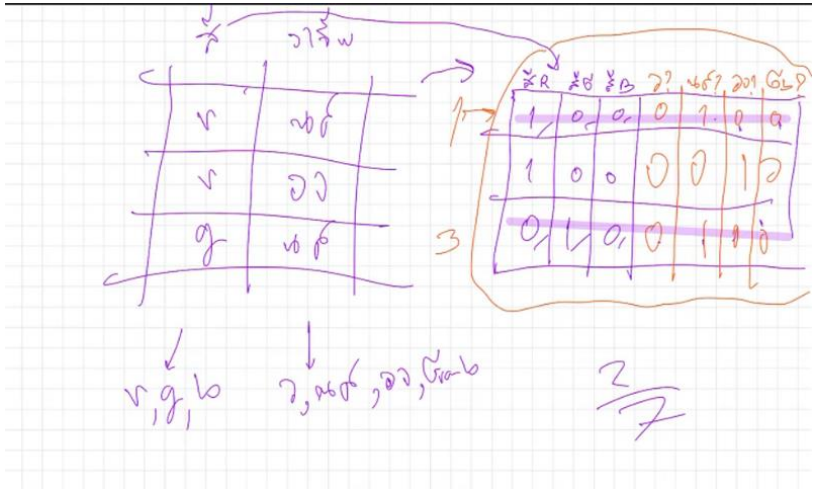
Proximity Measure for Categorical Attributes

- 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

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

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

- ระยะทางระหว่าง attributes categorical
- Categorical คือ เป็นชื่ออย่างเดียว
- วิธีการมี 2 แบบ แบบแรกคล้าย ๆ Similarity
- p– m คือจำนวนตัวที่ไม่เหมือน
- p ข้างล่างคือจำนวนทั้งหมด



- วิธีการในแบบที่ 2 ลักษณะตัวอย่าง

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: $r_{if} \in \{1,...,M_f\}$
  - Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - 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

- การหาระยะห่างระหว่างจุด ตามสูตร Z_{if}
-

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: $r_{if} \in \{1,...,M_f\}$
  - Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
    - 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

- ตัวอย่างการหาค่า Z_{if}

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

- เป็นวิธีการรวม เช่น ตัวอย่าง data โควิดหรือวงใน

- 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

- ตัวอย่างคำนวณวิธีการใช้งาน

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

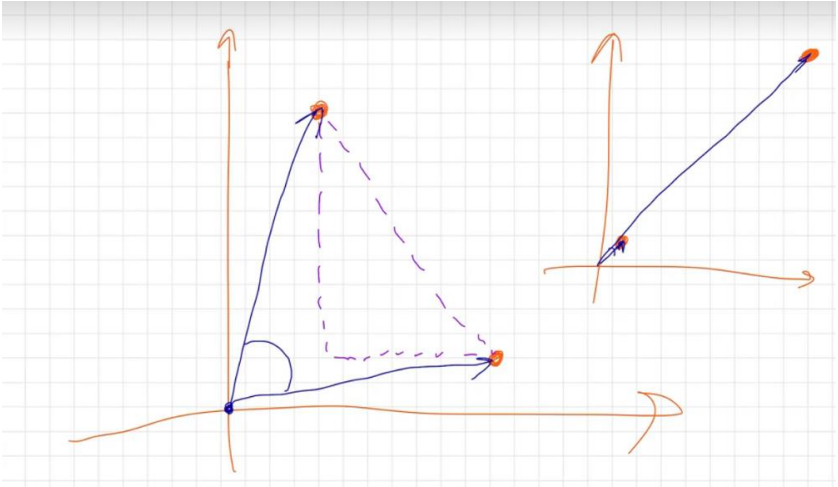
Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- 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*

- ตัววัดความเหมือนของจุด
- วัดความเหมือนหรือต่างกันยังไง



- ตัวอย่างให้เห็นภาพในการวัดระยะห่างระหว่างจุดอาจจะทำแบบการหามุมองศาแล้วหาระยะห่างก็ได้