Proximity Measure for Binary Attributes

- A contingency table for binary data

Binary มีได้แค่ 2 ค่า และแปลงข้อมูลแล้วมี
ค่าเป็น 0,1 เท่านั้น

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

- 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)

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

Distance measure for symmetric binary variable (มีความน่าจะเป็นหรือมีค่าเท่าๆกัน)

หาระยะห่างระหว่างจุด 2 จุด

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	Y	P	N	N	N	N

	1	0	sum
1	2 q	1 r	3
0	1 s	3 t	4
sum	3	4	7

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

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

$$= \frac{2}{7}$$

Distance measure for symmetric binary variable (มีความน่าจะเป็นที่จะมีค่าเท่าๆกัน)

หาระยะห่างระหว่างจุด 2 จุด

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	Y	N	P	N	P	N
Jim	M 1	Y 1	P 1	N 0	N 0	N 0	N 0

	1	0	sum
1	2 q	1 r	3
0	1 s	3 t	4
sum	3	4	7

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

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

$$= \frac{2}{7}$$

มีความนัยจะเป็นทั้งค่าที่ได้จะไม่เท่ากัน

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

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

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

Proximity Measure for Categorical Attributes

- Categorical data, also called nominal attributes เก็บเป็นชื่อ เช่น สี อย่างสีแดง สีเหลือง สีน้ำเงิน
  ↓ ตัวอักษร
  - Example: Color (red, yellow, blue, green), profession, etc.
- <u>Method 1</u>: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i,j) = \frac{\overset{\text{จำนวนทั้งหมด}}{p} - \overset{\text{จำนวนที่เหมือนกัน}}{m}}{\underset{\text{จำนวนทั้งหมด}}{p}}$$

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

สี → R, G, B   อาชีพ → ว่างงาน, อาจารย์, นักศึกษา, GRAB

สี	อาชีพ
R	น.ศ.
R	อ.
G	น.ศ.

→

R	G	B	ว่างงาน	อ.	น.ศ.	GRAB
1	0	0	0	0	1	0
1	0	0	0	1	0	0
0	1	0	0	0	1	0

$p = 7$

$m = 2$

$\frac{2}{7}$

Attributes of Mixed Type

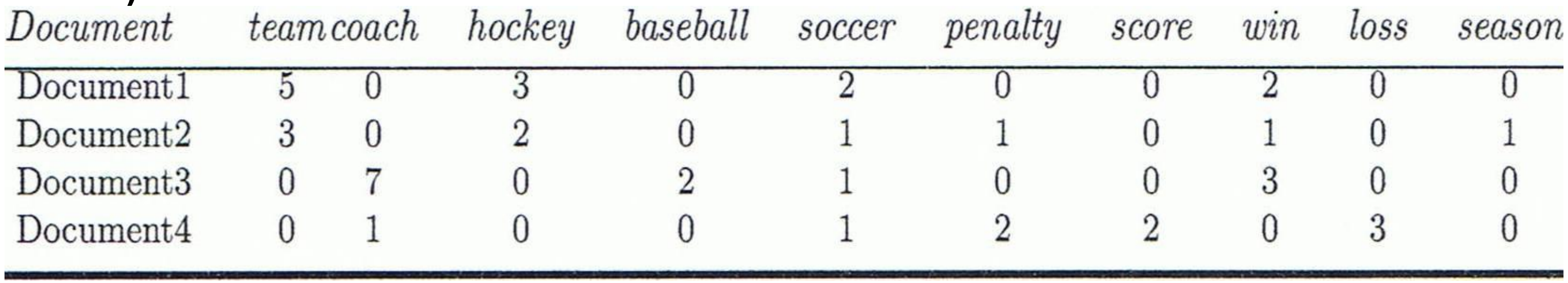
- A dataset may contain all attribute types ขอมูลสามารถเรียงลำดับได้
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

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

  - 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

- 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

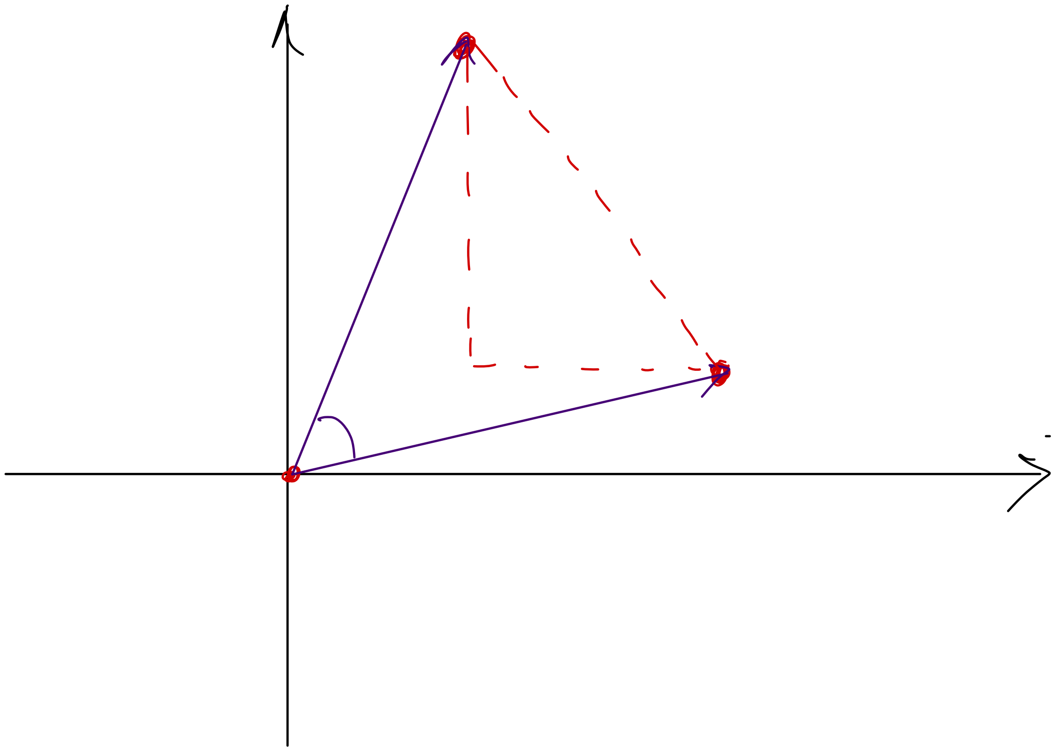


Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- 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

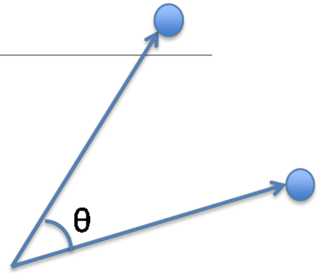


วัดความห่างด้วยมุม

- ถ้ามุมมีองศามาก = ข้อมูลทั้งสองต่างกันมาก
- ถ้ามุมมีองศาน้อย = ข้อมูลต่างกันน้อย

ใช้ได้กับข้อมูลที่มีความมากน้อยต่างกันได้ เนื่องจากใช้องศาในการวัด แตกต่างกัน

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||

$$\| d_1 \| = \sqrt{5\times5+0\times0+3\times3+0\times0+2\times2+0\times0+0\times0+2\times2+0\times0+0\times0} = 6.481$$

$$\| d_2 \| = \sqrt{3\times3+0\times0+2\times2+0\times0+1\times1+1\times1+0\times0+1\times1+0\times0+1\times1} = 4.12$$

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

Announcements: Meetine of the 4th Credit Project

- CS412: **Assignment #1** was distributed last Tuesday!
  - The due date is Sept. 15. No late homework will be accepted!!
- **Waitlist is cleared**:   We took 50 additional students into the video only session
  - Please find your status with Holly.   You are either in or out (wait for Spring 2017)
- Meeting for **Project for the 4th Credit**
  - You can change from 4 to 3 credit or from 3 to 4 credits by sending me e-mails
  - **Meeting time and location: 10-11am Friday (tomorrow!) at 0216 SC**
  - This project is part of WSDM 2017 Cup
  - Choice #1: **Triple Scoring**: Computing relevance scores for triples from type-like relations
  - Choice #2: **Vandalism Detection** for Wikipages
  - Tas/PhD student/postdoc will give you the details in the Friday meeting! **Must attend if you want to do the 4th credit project!!!**