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

Binary คือเป็นได้แค่ 2 ค่า และเมื่อแปลงข้อมูลแล้วข้อมูลจะมีค่าแค่ค่า 0,1 เท่านั้น

- 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)

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

62

Symmetric Binary Variables (ค่าความน่าจะเป็นที่จะมีค่าเท่าๆกัน)

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

Jack

Mary \	1	0	Sum
1	2 (q)	1 (r)	3
0	1 (s)	3 (t)	4
Sum	1	3	7

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

1+1/2+1+1+3 = 2/7

มีทั้งหมด 7 ค่า

ต่างกัน 2ค่า

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

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

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

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

Jack \ Mary

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

Jack \ Jim

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

Jim \ Mary

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

Dummy Variables

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

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

จำนวนทั้งหมด - ตัวที่เหมือนกัน

จำนวนทั้งหมด

เป็นชื่อ ที่อยู่ในประเภท เช่น อาชีพ ก็จะมี หมอ พยาบาล อาจารย์ เป็นต้น

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

**Sklearn.preprocessing.OneHotEncoder

64

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

64

สี -> R,G,B อาชีพ -> ว่างงาน, อาจารย์, นักศึกษา, Grab

สี	อาชีพ
R	นักศึกษา
R	อาจารย์
G	นักศึกษา

ทำการแปลงค่าก่อน

สี R	สี G	สี B	ว่างงาน	อาจารย์	นักศึกษา	Grab
1	0	0	0	0	1	0
1	0	0	0	1	0	0
0	1	0	0	0	1	0

จุดที่ 1 กับ 3 ห่างกัน 2/7

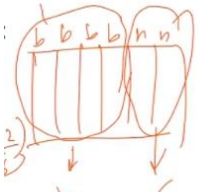
Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank (e.g., freshman, sophomore, junior, senior) สามารถเรียงลำดับได้
  1 2 3 4
- 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

Attributes of Mixed Type

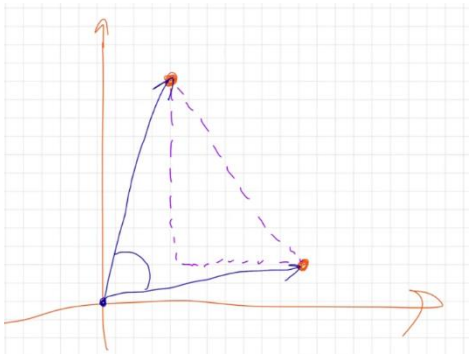
- A dataset may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, and ordinal
- One may use a weighted formula to combine their effects:

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



Y(4/6)+X(2/6)

  - 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



วัดความห่างด้วยมุม

ถ้ามุมมีองศาที่มาก หมายความว่า ของมูลทั้งสองตัวมีความต่างกันมาก

ถ้ามุมมีองศาที่น้อย หมายความว่า ข้อมูลทั้งสองมีความแตกต่างกันน้อย

ใช้ได้กับข้อมูลที่มีความมากน้อยต่างกันได้ เนื่องจากใช้องศาในการวัดความแตกต่าง

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

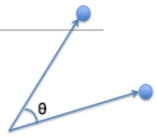
- Other vector objects: Gene features in micro-arrays
- Applications: Information retrieval, biologic taxonomy, gene feature mapping, etc.
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

where • indicates vector dot product, $||d||$: the length of vector d

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