Measure of Similarity and Dissimilarity

Lecture 4.3
28 Jul 2020

Recap

- Started with concepts of Data Mining
  - Techniques
  - Challenges
- Reviewed Machine Learning concepts
  - Types
  - Terms
- Measuring similarity and Dissimilarity
- Concept of data matrix and distance/dissimilarity matrix

Recap

- Measuring similarity and dissimilarity
  - Nominal attributes → categorical
  - Binary attributes → Symmetric and asymmetric attributes
  - Numeric attributes
  - Ordinal attributes

Recap

- For numeric attributes
  - Euclidean Distance
  - Manhattan Distance
  - Minkowski Distance
    - $$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h},$$

Today’s plan

- Measuring Proximity for mixed type attributes
- Cosine similarity

Proximity measure for mixed attributes

- In reality, objects are represented by mixture of attributes
- How to compute similarity/dissimilarity?
- One approach
  - Group (cluster) attributes by type
  - Compute the dissimilarity for each group
  - Compare the similarity

Proximity measure for mixed attributes

- In reality, objects are represented by mixture of attributes
- How to compute similarity/dissimilarity?
- One approach
  - Group (cluster) attributes by type
  - Compute the dissimilarity for each group
  - Compare the similarity

This will not work in all cases

Proximity measure for mixed attribute

- More preferable approach is
  - Process all attribute types together
  - Prepare a single dissimilarity matrix
  - Bring all attributes into a common scale of [0,1]

Proximity measure for mixed attribute

- If dataset contain p attributes of mixed types
- Then we compute d(i,j) as:

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

Proximity measure for mixed attributes

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

where the indicator $\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing (i.e., there is no measurement of attribute f for object i or object j), or (2) $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary; otherwise, $\delta_{ij}^{(f)} = 1$. The contribution of attribute f to the dissimilarity between i and j (i.e., $d_{ij}^{(f)}$) is computed dependent on its type:

- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f.
- If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as numeric.

Example

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Compute distance matrix for each attribute separately
- Then compute dissimilarity matrix for mixed attributes (for the dataset
- Apply

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

Example

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Test -1 is nominal
- For nominal $d_{ij}^{(f)}$ = 0 if attributes x_{if} and x_{jf} are same , 1 otherwise
-

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

Example

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Test -2 is ordinal
- First compute rank for attribute r_{if}
- Compute $Z_{if} = (r_{if} - 1)/(M_f-1)$
- Then treat Z_{if} as numeric (find the Euclidean distance)

Value	Rank	Computation	Z_{if}
Excellent	1	(1-1)/(3-1)	0
Good	2	(2-1)/(3-1)	0.5
Fair	3	(3-1)/(3-1)	1

	1	2	3	4
1	0			
2	1.0	0		
3	0.5	0.5	0	
4	0	1.0	0.5	0

Example

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Test -3 is numeric

If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$,

- Here, max = 64, min 22

	1	2	3	4
1	0			
2	0.55	0		
3	0.45	1.0	0	
4	0.4	0.14	0.86	0.

Example

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

Nominal

	1	2	3	4
1	0			
2	1.0	0		
3	0.5	0.5	0	
4	0	1.0	0.5	0

Ordinal

	1	2	3	4
1	0			
2	0.55	0		
3	0.45	1.0	0	
4	0.4	0.14	0.86	0.

Numeric

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

d(2,1) = (1*1 + 1*1 + 1*0.55) / (1+1+1)
= 2.55/3
= 0.85

Example

Complete the matrix

Then say

Which are the most similar data items?

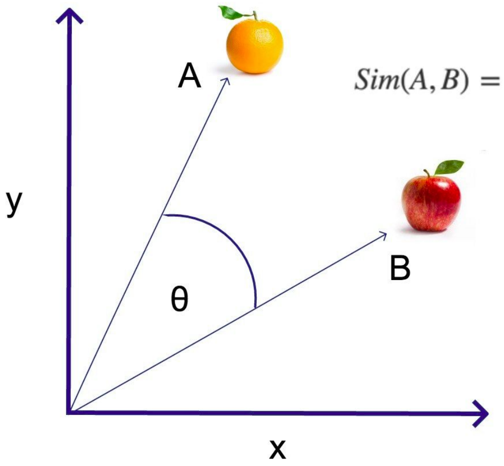
Which are least similar?

Cosine similarity

- Vector similarity calculation
- Document -Document similarity
- Document-Query similarity

- Cos (0) = 1
- Cos (90) = 0
- Cos (180) = -1

Cosine Similarity



$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

Example

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

x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$$\boldsymbol{x}^t \cdot \boldsymbol{y} = 5\times 3 + 0\times 0 + 3\times 2 + 0\times 0 + 2\times 1 + 0\times 1 + 0\times 0 + 2\times 1 + 0\times 0 + 0\times 1 = 25$$

$$||\boldsymbol{x}|| = \sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2} = 6.48$$

$$||\boldsymbol{y}|| = \sqrt{3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2} = 4.12$$

$$sim(\boldsymbol{x}, \boldsymbol{y}) = 0.94$$

Summary

- Similarity , Dissimilarity (or distance) , Proximity
- Measuring proximity for different attributes
- Cosine similarity

Next

- Review Probability
- Important distributions used in MLDM

Announcements

- There is a change in slot, **only for next week**
  - **04 Aug 2020 - Tuesday - Slot 1 - 15CSE376 - NCP - Dr. Arunkumar C**
  - **07 Aug 2020 - Friday - Slot-1 - 15CSE401 - MLDM -Manu Madhavan**

- **Today's** (28 Jul 2020) discussion slot will be from 04-05 PM (instead of 06-07 PM).

- **Form your group** (4 per group) for case study, think about the topic.

- Gentle reminder for **Tomorrow's (29 Jul 2020) quiz**.

Thank you