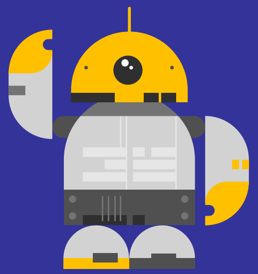


Database Systems

Lecture 6 Cont. - Getting to Know Your Data

Dr. T. Akilan

takilan@lakeheadu.ca

This Session

- Measuring Data Similarity and Dissimilarity

Data Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity (e.g., distance)
  - Numerical measure of how different two data objects are
  - Value is lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- "Proximity" can refer to similarity or dissimilarity

Data Matrix vs Dissimilarity Matrix

- Assume a set of n samples
  - Each sample has a set of p attributes or features
- Two data structures:
  - Data matrix (object-by-attribute structure)
  - Dissimilarity matrix (object-by-object structure)
- Data matrix: n-by-p matrix
  - form of a relational table with n objects $\times$ p attributes
- Dissimilarity matrix: n-by-n table
  - Difference between samples i and j, $d(i,j)$
    - ✓ We need to define a way to compute it
  - Similarity $= 1 - d(i,j)$

feature vectors

$$X = \{x_1, x_2, \ldots, x_N\}$$

$$x_1 = (x_{11}, x_{12}, \ldots, x_{1p})$$

$$x_2 = (x_{21}, x_{22}, \ldots, x_{2p})$$

$$\vdots$$

two-mode matrix

$$X = \begin{bmatrix} x_{1,0} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,0} & \cdots & x_{n,p} \end{bmatrix}$$

one-mode matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Dissimilarity of Nominal Attributes

- Descriptive (qualitative) attribute with no inherent quantitative value
  - No order in the values, no way to measure the level of difference between them
  - Only measure is "are they the same or not?"
  - Can be computed based on the ratio of mismatches:

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

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

A Sample Data Table Containing Attributes of Mixed Type

- Binary dissimilarity matrix:
  - Only one nominal attribute (e.g., test-1)
  - 1 – objects' attributes are different
  - 0 - Objects' attributes are the same

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

→ Multiple multi-level nominal attributes

Dissimilarity of Nominal Attributes Cont.

- Dissimilarity matrix for objects with more than one **multi-level nominal** attribute

Car	Maker	Model	Colour
Car 0	Acura	TSX	Silver
Car 1	VW	Beetle	White
Car 2	Ford	Model T	Silver
Car 3	Acura	TSX	Black

$$d(i,j) = \frac{P - m_{i,j}}{P}$$

Example:

$$d(car_0, car_1) = \frac{3-0}{3} = 1,$$

$$d(car_0, car_2) = \frac{3-(0+0+1)}{3} = \frac{2}{3} = 0.67$$

$$d = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.67 & 1 & 0 & \\ 0.33 & 1 & 1 & 0 \end{bmatrix}$$

- Alternatively, similarity can be computed as: $sim(i,j) = 1 - d(i,j) = \frac{m}{p}$

Dissimilarity of Binary Attributes

- Recall that a binary attribute has only one of two states: 0 and 1, where 0 means that the attribute is absent, and 1 means that it is present
  - E.g., the attribute smoker describing a patient, for instance, 1 indicates that the patient smokes, while 0 indicates that the patient does not.

- Use 2×2 contingency table:

		Object ***j***	
		1	**0**
Object *i*	**1**	**PP**	**PN**
	0	**NP**	**NN**

  - PP - # of attributes that equal 1 for both objects i and j,
  - PN - # of attributes that equal 1 for object i but equal 0 for object j,
  - NP - # of attributes that equal 0 for object i but equal 1 for object j,
  - NN - # of attributes that equal 0 for both objects i and j.

Dissimilarity of Binary Attributes Cont.

- Symmetric attributes dissimilarity:
  - Both states are equally important

$$d(i, j) = \frac{PN + NP}{PP + PN + NP + NN}$$

		Object *j* 1	Object *j* 0
Object *i*	1	PP	PN
	0	NP	NN

- Asymmetric attributes dissimilarity:
  - Positives are much more important
  - We don't really care for 0-0 matches

$$d(i, j) = \frac{PN + NP}{PP + PN + NP}$$

- Jaccard coefficient is asymmetric similarity

$$sim(i, j) = 1 - d(i, j) = \frac{PP}{PP + PN + NP}$$

Dissimilarity of Binary Attributes – Example

- A relational table, where patients are described by a set of binary attributes: name, gender, fever, cough, test-1, test-2, test-3, and test-4

Object identifier Symmetric Asymmetric

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Dissimilarity of Binary Attributes – Example

- Compute patients' **dissimilarity** based only on the **asymmetric attributes**.

Symmetric: gender; Asymmetric: fever, cough, test-1, test-2, test-3, test-4

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$d(i, j) = \frac{PN + NP}{PP + PN + NP}$$

		Object j	
		1	**0**
Object i	**1**	**PP**	**PN**
	0	**NP**	**NN**

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67,$$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33,$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75.$$

- What can you suggest from these measures?

Dissimilarity of Numeric Attributes

- Attributes have quantitative values
  - o We are not limited to "are they the same"
  - o We can meaningfully talk about "how different are they"

  Let $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects

- Euclidian (straight-line) distance: $d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$

- Manhattan (city-block) distance: $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$

- Minkowski distance: $d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$

- Supremum (L_{max}, L_∞ norm, Chebyshev) distance: $d(i, j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$

Dissimilarity of Numeric Attributes – Working Example 2:00

- E.g., two objects represented by two attributes: $x_1 = (1,2)$ and $x_2 = (3,5)$

Euclidean distance: $d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$

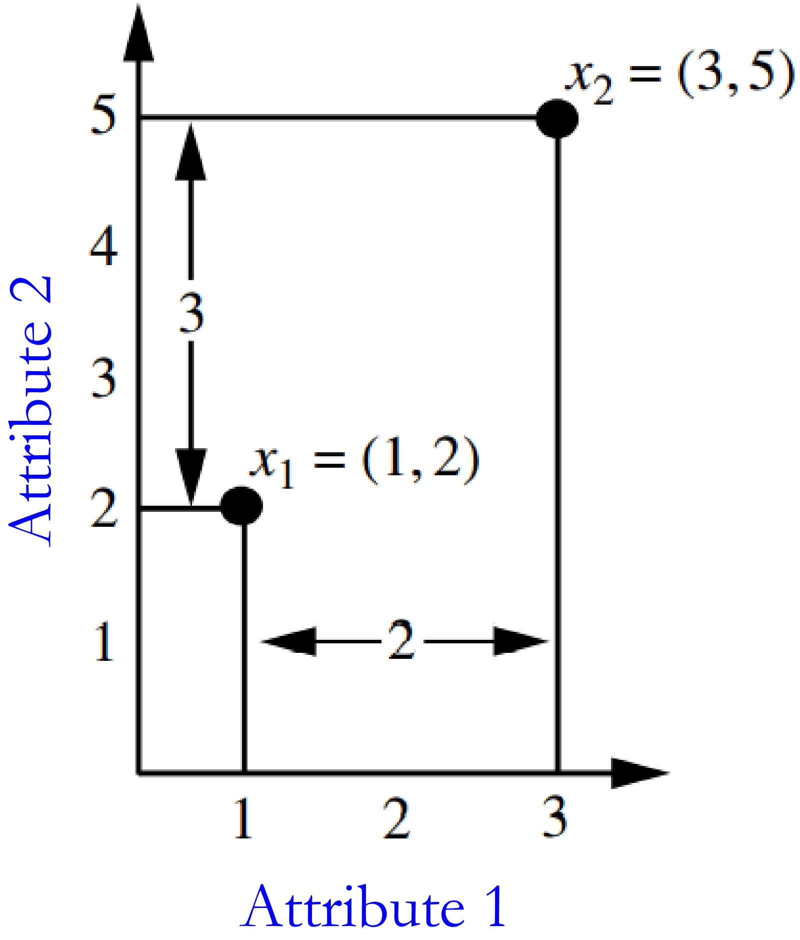
$$d(i,j) = \sqrt{(1-3)^2 + (2-5)^2} = 3.61$$

Manhattan distance: $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$

$$d(i,j) = |1-3| + |2-5| = 5$$

Supremum distance: $d(i,j) = \max_{f}^{p} |x_{if} - x_{jf}|$

$$d(i,j) = \max(|1-3|, |2-5|) = 3$$



Dissimilarity of Ordinal Attributes

- Ranked & ordered categories
  - {"cold", "warm", "hot"}
  - "cold" is more similar to "warm" than to "hot"
- Magnitude between successive ranks is unknown
  - How many degrees of difference between "cold" and "warm"?
- Solution: Number the ranks
  - {"cold", "warm", "hot"} = {0, 0.5, 1}
    - ✓ Using normalized numbers in [0.0, 1.0] avoids the problem of more detailed sets looking more significant
      - ➢ {"cold", "lukewarm", "warm", "hot", "crazy hot"} as {0.0, 0.25, 0.5, 0.75, 1.0} puts in in the same range as the other set
      - ➢ {0, 1, 2, 3, 4} makes the top value look 4 times as important as the top of the other set, which is wrong
  - Then use one of the **distance measures from numeric attributes**

Dissimilarity of Ordinal Attributes – Example

- Build the dissimilarity matrix
  - User rating scale: {fair, good, excellent}
  - CAA rating scale: {*, **, ***, ****}
  - Use Euclidian distance

Car	User rating	CAA rating
Car 0	excellent	**
Car 1	fair	**
Car 2	good	***
Car 3	excellent	****

- **Step 1** - Replace each attribute values by rank and normalize the rank.

Object	User rating			CAA rating		
		r	nd		r	nd
Car 0	E	3	1	**	2	0.33
Car 1	F	1	0	**	2	0.33
Car 2	G	2	0.5	***	3	0.67
Car 3	E	3	1	****	4	1

{fair, good, excellent} → {0, 0.5, 1}
{*, **, ***, ****} → {0, 0.33, 0.67, 1}

- **Step 2** – Use a numeric attribute's dissimilarity measure. In this case, Euclidian distance.

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.60 & 0.60 & 0 & \\ 0.67 & 1.20 & 0.60 & 0 \end{bmatrix}$$

Dissimilarity of Mixed Attributes

- Objects can have properties of several different types

- We need to compare and combine the differences in all their values

- Difference between two objects *i* and *j* that have *P* attributes: 1, 2,…, *f*,…, *P*

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

- $d^{(f)}$ - dissimilarity value of attribute *f*
- $\delta^{(f)}$ determines if attribute *f* should be counted, i.e., whether it contributes in the distance measure.

Let's see how delta determines the featurecontribution

Dissimilarity of Mixed Attributes

- $\delta^{(f)}$ determines if attribute f should be counted

$$\delta_{ij}^{(f)} = \begin{cases} 0 & \text{if } x_{if} \text{ or } x_{jf} \text{ is missing} \\ 0 & x_{if} = x_{jf} = 0 \text{ and } f \text{ is asymmetric binary} \\ 1 & \text{otherwise} \end{cases}$$

- $d^{(f)}$ is the dissimilarity value of attribute f
    - To be fair, all attributes need to be normalized to [0,1]
    - Already the case for nominal, binary and ordinal
    - For numeric attributes, normalize by dividing by the maximum distance between two samples:

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

: h runs over all non-missing objects for attribute f.

Dissimilarity of Mixed Attributes – Example

- Build the dissimilarity matrix
  - Price is numeric (use Manhattan), Topic is nominal, Rating is ordinal

Book ID	Price	Topic	Rating
0	45	Arts	Excellent
1	22	Business	Fair
2	64	Engineering	Good
3	28	Arts	Excellent

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

- Apply specific distance measure for each attribute: $d(f)$

d(1, 0)

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

d(3, 0)

Document Similarity - Cosine Distance

- Good for dealing with sparse matrix of attributes
  - Large matrix with a lot of zeros
- Example: term-frequency vector
  - Typical in Natural Language Processing
  - Represent text document by counting words

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

Document Vector or Term-Frequency Vector

Cosine Similarity

- Cosine measure as a similarity function: $sim(\boldsymbol{x}, \boldsymbol{y}) = \dfrac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||}$

$$\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$

- E.g.,:

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

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

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

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

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

- **How similar are Doc 1 and Doc2?**
- Let x and y represent tf-vector of Doc 1 and Doc2 respectively.

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

Summary

- A data object is composed of attributes that have values

- The attributes can be nominal, binary (symmetric or asymmetric), ordinal, numeric (interval-scaled or ratio-scaled)

- We can analyze the data in terms of its attributes' values
  - Central tendency: mean, median, mode
  - Dispersion: range, quantiles, variance, standard deviation
  - Redundancy: χ^2 correlation, correlation, covariance

- We can measure the difference and similarity between pairs of objects based on their attributes' values

References

[1] Jacqueline S. McLaughlin at The Pennsylvania State University. In turn citing: R. A. Fisher and F. Yates, Statistical Tables for Biological Agricultural and Medical Research, 6th ed

[2] Jiawei Han, Micheline Kamber, & Jian Pei, *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, ISBN: 978-0-12-381479-1