Measures of Similarity and Dissimilarity

Lecture 4.2
25 Jul 2020

Recap

- Concept of similarity and dissimilarity (distance)
- Proximity
- Use of similarity measures in data mining and machine learning
- Types of Attributes
  - Nominal
  - Binary
  - Ordinal
  - Numeric
- Proximity measure of nominal attributes
  - Simple mismatch calculation

Today

- Proximity measures for
  - Binary
  - Numeric Data
    - Distance measures
      - Euclidean
      - Manhattan
      - Minkowski

Proximity measures for Binary attributes

- Only two state: 0 (absent) and 1 (present)
- Binary attribute can be:
  - **Symmetric** : if **both of its states are equally valuable**

  - **Asymmetric**: if the outcome of the states are not equally important

Proximity measures for Binary attributes

- Only two state: 0 (absent) and 1 (present)
- Binary attribute can be:
  - **Symmetric** : if **both of its states are equally valuable**
  - **Asymmetric**: if the outcome of the states are not equally important

Think examples

Proximity measures for Binary attributes

- Only two state: 0 (absent) and 1 (present)
- Binary attribute can be:
  - **Symmetric** : if **both of its states are equally valuable**
    Gender (male/female)
  - **Asymmetric**: if the outcome of the states are not equally important
    Disease case (Corona Positive is more important than negative)
    Agreement of two 1's is more important than two 0's

Contingency matrix

Contingency Table for Binary Attributes

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

- Let object i and j represented by p binary attributes
- q = count (attributes 1 for both objects)
- r = count (attributes 1 for i and 0 for j)
- s = count (attributes 0 for i and 1 for j)
- t = count (attributes 0 for both i and j)
- p = q+r+s+t

Binary Symmetric Attributes

- dis(i,j) = (r+s)/ (p) = (r+s)/ (q+r+s+t)
- What will be the similarity?

Contingency Table for Binary Attributes

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

Asymmetric Binary Attributes

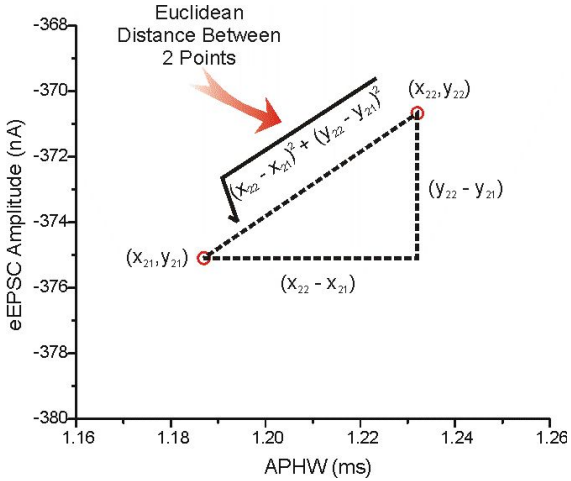
- Agreement of two 1’s is more important
- dis(i,j) = (r+s)/ **?**

Asymmetric Binary Attributes

- Agreement of two 1’s is more important
- dis(i,j) = (r+s)/ **(q+r+s)**
- Sim (i,j) = 1- dis(i,j) = (q)/(q+r+s)
  - **Jaccard coefficient**

Dissimilarity of Numeric Data

- Dissimilarity is generally measured by “Distance”
- Normalize data [optional]
- **Euclidean Distance:**
  -



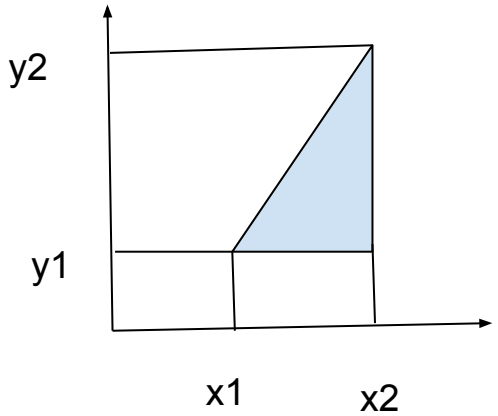
$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Dissimilarity of Numeric Data

Manhattan Distance

- $d(i,j) = |x_1-y_1| + |x_2-y_2| + ...+ |x_n - y_n|$



Try yourself

X = [1, 2, 4, 5]

Y = [2, 3, 3, 6]

Find Euclidean and Manhattan distance between X and Y?

Note: Normalize $x_i = (x_i - \min)/(\max-\min)$

You may write a code and try (optional)

Properties of Distance measures

- **Non negativity**
  - d(i,j) >= 0
- **Identity of indiscernible**
  - d(i,i) = 0
- **Symmetry**
  - d(i,j) =d(j,i)
- **Triangle inequality**
  - d(i,j) <= d(i,k) + d(k,j)
  - distance from i to j via k is at least as great as from i to j directly

Proximity measure for ordinal attribute

- For each ordinal attribute
  - Convert ordinal values to their rank
  - Apply normalization
- Then treat as numeric attribute and compute distance

Next

- Proximity measure for mixed type
  - Read 2.4.6 of Han, Data Mining Text book
- Cosine similarity
  - Read 2. 4.7 of Han, Data Mining Text Book

Announcement

- There will be a Quiz on **Wednesday, July 29 2020**
    - Topics: Data mining concepts, challenges, Machine Learning Recap
    - Refer Lectures 1,2,3 and Reading assignments
    - Quiz will be for 15 marks
- Please share this information to your class group