BITS Pilani

BITS Pilani
Hyderabad Campus

Prof.Aruna Malapati
Department of CSIS

innovate achieve lead

BITS Pilani
Hyderabad Campus

Similarity and Distance Measures

Today’s Learning objective

- What is Distance?
- Similarity vs. distance
- Properties of distance metrics
- Similarity Measures for Binary Nominal attributes
- Similarity Measures for Categorical Attribute
- Proximity Measures for Ordinal Attribute

What is Distance?

Let $\mathcal{S}$ be a space of data objects. A distance function has the type

$$d : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^{+} \cup \{0\}$$

Ituitively: Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}$ be objects.

- if $d(\mathbf{x}, \mathbf{y})$ small, $\mathbf{x}$ and $\mathbf{y}$ are close or similar
- If $d(\mathbf{x}, \mathbf{y}) < d(\mathbf{x}, \mathbf{z})$, $\mathbf{x}$ is closer/more similar to $\mathbf{y}$ than $\mathbf{z}$

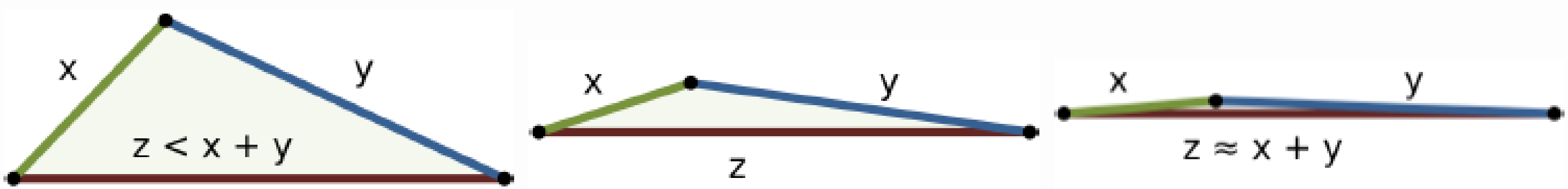
Similarity vs. distance

Similarity function $s : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$

- $s(\mathbf{x}, \mathbf{y})$ large when $\mathbf{x}$ and $\mathbf{y}$ similar (and $d(\mathbf{x}, \mathbf{y})$ small)
- often $s : \mathcal{S} \times \mathcal{S} \to [0, 1]$
- $\Rightarrow$ possible to induce distance $d_s = 1 - s$
- if $d : \mathcal{S} \times \mathcal{S} \to [0, 1]$, possible to induce similarity $s_d = 1 - d$
- if not, then e.g.,
  $s_d = 1 - \frac{d}{D}$ (D=maximal possible distance) or
  $s_d = \frac{1}{1+d}$

Metric: distance d *that satisfies 4 properties*

1. $d(\mathbf{x},\mathbf{y}) \geq 0$ (non-negativity or separation)
2. $d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (coincidence axiom)
3. $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ (symmetry)
4. $d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z})$ (**triangle inequality**)



Proximity

- Examples:
  - ✓ For an item bought by a customer, find other similar items
  - ✓ Group together the customers of the site so that similar customers are shown the same ad.
  - ✓ Group together web documents so that you can separate the ones that talk about politics and the ones that talk about sports.
  - ✓ Find all the near-duplicate mirrored web documents.
  - ✓ Find credit card transactions that are very different from previous transactions.
- To solve these problems, we need a definition of similarity or distance.
- For many problems, we need to quantify how close two objects are.

Similarity / Distance

Attribute Type	Similarity	Dissimilarity
Nominal	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
Ordinal	$s = 1 - \frac{\lVert p - q \rVert}{n - 1}$	$d = \frac{\lVert p - q \rVert}{n - 1}$
	(values mapped to integer 0 to n-1, where n is the number of values)	
Interval or Ratio	$s = 1 - \lVert p - q \rVert$, $s = \frac{1}{1 + \lVert p - q \rVert}$	$d = \lVert p - q \rVert$

Similarity Measures for Binary attribute

- ✓ Suppose a binary attribute Gender = {Male, female} where Male is equivalent to binary 1 and female is equivalent to binary 0.
- ✓ The similarity value(p) is 1 if the two objects contain the same attribute value, 0 otherwise.

Object	Gender
Ram	Male
Sita	Female
Laxman	Male

- ✓ $p(Ram, sita) = 0$
- ✓ $p(Ram, Laxman) = 1$

- ✓ Note : In this case, if q denotes the dissimilarity between two objects i *and* j with single binary attributes, then $q_{(i,j)} = 1 - p_{(i,j)}$

Proximity Measures for Two or more Binary attribute

✓ We define the contingency table summarizing the different matches and mismatches between any two objects x and y, which are as follows.

Contingency table with binary attributes

		Object y	
Object x		1	0
	1	f_{11}	f_{10}
	0	f_{01}	f_{00}

Here, f_{11}= the number of attributes where x=1 and y=1.

f_{10}= the number of attributes where x=1 and y=0.

f_{01}= the number of attributes where x=0 and y=1.

Similarity Measure for Symmetric Binary attribute

- Symmetric binary coefficient($\mathcal{S}$) is used to measure the similarity between two objects and is defined as

$$\mathcal{S} = \frac{Number\ of\ matching\ attribute\ values}{Total\ number\ of\ attributes}$$ or

$$\mathcal{S} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Similarity Measure with Symmetric Binary

Consider the following two dataset, where objects are defined with symmetric binary attributes.

Gender = {M, F}, Food = {V, N}, Caste = {H, M},Education = {L, I}, Hobby = {T, C},Job = {Y, N}

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

	1	0
1	1	1
0	2	2

$$\mathcal{S}(\text{Hari, Ram}) = \frac{2+1}{2+2+1+1} = 0.5$$

Proximity Measure with Asymmetric Binary

- Jaccard Coefficient is used to measure the similarity between two objects is symbolized by $\mathcal{J}$ and is defined as follows

$$\mathcal{J} = \frac{Number\ of\ matching\ presence}{Number\ of\ attributes\ not\ involved\ in\ 00\ matching}$$

or

$$\mathcal{J} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Proximity Measure with Asymmetric Binary

Consider the following two dataset.

Gender = {M, F}, Food = {V, N}, Caste = {H, M}, Education = {L, I},Hobby = {T, C}, Job = {Y, N}

Compute the Jaccard coefficient between Ram and Hari assuming that all binary attributes are asymmetric and for each pair values for an attribute, first one is more important than the second.

Object	Gender	Food	Caste	Education	Hobby	Job
Hari	M	V	M	L	C	N
Ram	M	N	M	I	T	N
Tomi	F	N	H	L	C	Y

	1	0
1	1	1
0	2	2

$$\mathcal{J}(\text{Hari, Ram}) = \frac{1}{2+1+1} = 0.25$$

Note: $\mathcal{J}(\text{Hari, Ram}) = \mathcal{J}(\text{Ram, Hari})$

Proximity Measures for Categorical Attribute

- Attributes with three or more states (e.g. color = {Red, Green, Blue}) are called nominal.
- If $s(x, y)$ denotes the similarity between two objects $x \text{ and } y$, then

$$s(x,y) = \frac{\text{Number of matches}}{\text{Total number of attributes}}$$

- and the dissimilarity $d(x, y)$ is

$$d(x,y) = \frac{\text{Number of mismatches}}{\text{Total number of attributes}}$$

- If m = number of matches and a = number of the categorical attribute for object x and y then s and D are defined as

$$s(x,y) = \frac{m}{a} \quad \text{and} \quad d(x,y) = \frac{a-m}{a}$$

Proximity Measures for Categorical Attribute

Object	Color	Position	Distance
1	R	L	L
2	B	C	M
3	G	R	M
4	R	L	H

Proximity Measure for Ordinal Attribute

- Ordinal attribute is a special kind of categorical attribute, where the values of attribute follow a sequence (ordering), e.g., Grade = {Ex, A, B, C} where Ex > A >B >C.
- Suppose, *A* is an attribute of type ordinal and the set of values of $A = \{a_1, a_2, \ldots\ldots, a_n\}$. Let n values of A are ordered in ascending order as $a_1 < a_2 < \ldots < a_n$. Let *i-th* attribute value a_i be ranked as *i, i=1,2,..n.*
- The normalized value of a_i can be expressed as

$$\hat{a}_i = \frac{i-1}{n-1}$$

- Thus, normalized values lie in the range $[0..1]$.
- As a_i is a numerical value, the similarity measure, then can be calculated using any similarity measurement method for numerical attribute.
- For example, the similarity measure between two objects $x \text{ and } y$ with attribute values a_i and a_j, then can be expressed as

$$s(x, y) = \sqrt{(\hat{a}_i - \hat{a}_j)^2}$$

where $\hat{a}_i$ and $\hat{a}_i$ are the normalized values of $\hat{a}_i$ and $\hat{a}_i$, respectively.

Proximity Measure for Ordinal Attribute

Consider the following set of records, where each record is defined by two ordinal attributes size={S, M, L} and Quality = {Ex, A, B, C} such that S<M<L and Ex>A>B>C.

Object	Size	Quality
A	S (0.0)	A (0.66)
B	L (1.0)	Ex (1.0)
C	L (1.0)	C (0.0)
D	M (0.5)	B (0.33)

S=1=1-1/3-1=0
M=2=2-1/3-1=0.5
L=3=3-1/3-1=1

A=1= 1-1/4-1=0
B=2=2-1/4-1=0.33
C=3=3-1/4-1= 0.66
Ex=4 = 4-1/4-1 = 1

Proximity Measure with Interval Scale

- The generic formula to express distance *d* between two objects x and y in *n*-dimensional space.

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{\frac{1}{r}}$$

Here, r is any integer value, x_i and y_i denote the values of i^{th} attribute of the objects x and y respectively

- This distance metric most popularly known as Minkowski metric.

Proximity Measure with Interval Scale

Manhattan distance (L_1 Norm: r = 1)

The Manhattan distance is expressed as

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

where |…| denotes the absolute value.

This metric is also alternatively termed as **Taxicabs metric, city-block metric**.

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance is $|7-3| + |3-2| + |5-6| = 6.$

As a special instance of Manhattan distance, when attribute values $\in [0,1]$ is called Hamming distance.

Alternatively, Hamming distance is the number of bits that are different

Proximity Measure with Interval Scale

Euclidean Distance (L_2 Norm: r = 2)

This metric is same as Euclidean distance between any two points $x \text{ and } y \text{ in } \mathcal{R}^n$.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

Proximity Measure with Interval Scale

Chebychev Distance (L$_\propto$ Norm: $r \in \mathcal{R}$)

This metric is defined as

$$d(x, y) = \max_{\forall i}\{|x_i - y_i|\}$$

Example: x = [7, 3, 5] and y = [3, 2, 6].

The Manhattan distance = $|7 - 3| + |3 - 2| + |5 - 6| = 6.$

The chebychev distance = $\text{Max}\ \{|7 - 3|, |3 - 2|, |5 - 6|\} = 4.$

Proximity Measure for Ratio scale

The proximity between the objects with ratio-scaled variable can be carried with the following steps:

1. Apply appropriate transformation to the data to bring it into a linear scale. (e.g. logarithmic transformation to data of the form $X = Ae^{B}$.
2. The transformed values can be treated as interval-scaled values. Any distance measure discussed for interval-scaled variable can be applied to measure the similarity.

Proximity Measure for Ratio scale

Normalization:

- A major problem when using the similarity (or dissimilarity) measures (such as Euclidean distance) is that the large values frequently swamp the small ones.
- For example, consider the following data.

Make	Cost 1	Cost 2	Cost 3
X	2,00,000	70	10
Y	2,50,000	100	5

- Here, the contribution of Cost 2 and Cost 3 is insignificant compared to Cost 1 so far the Euclidean distance is concerned.
- This problem can be avoided if we consider the normalized values of all numerical attributes.

Proximity Measure for Mixed Attributes

- The previous metrics on similarity measures assume that all the attributes were of the same type. Thus, a general approach is needed when the attributes are of different types.
- One straightforward approach is to compute the similarity between each attribute separately and then combine these attribute using a method that results in a similarity between 0 and 1.
- Typically, the overall similarity is defined as the average of all the individual attribute similarities.

Proximity Measure with Mixed Attributes

Object	A (Binary)	B (Categorical)	C (Ordinal)	D (Numeric)	E (Numeric)
1	Y	R	X	475	10^8
2	N	R	A	10	10^{-2}
3	N	B	C	1000	10^5
4	Y	G	B	500	10^3
5	Y	B	A	80	1

Take Home message

- Many algorithms compute proximity using either similarity or dissimilarity.
- The distance metric used will depend on the type of Feature/attribute.