Data Mining: Data

Lecture Notes for Chapter 2

Introduction to Data Mining , 2nd Edition

by

Tan, Steinbach, Kumar

Outline

- Similarity and Distance

Introduction

- Similarity and dissimilarity are important because they are used by a number of data mining techniques, such as clustering, nearest neighbor classification, and anomaly detection.
- In many cases, the initial data set is not needed once these similarities or dissimilarities have been computed. Such approaches can be viewed as transforming the data to a similarity (dissimilarity) space and then performing the analysis.

Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
  - The term **distance** is used as a synonym for dissimilarity; however, distance often refers to a special class of dissimilarities
- Proximity refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d,\ s = \frac{1}{1+d},\ s = e^{-d},$ $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

s = 0, 1, 10, 100 → 0, -1, -10, -100 [s = -d] (![0,1])
s = 0, 1, 10, 100 → 1, 0.5, 0.09, 0.01 [s = 1/(1+d)]
s = 0, 1, 10, 100 → 1, 0.37, 0, 0 [s = e^{-d}]
s = 0, 1, 10, 100 → 1, 0.99, 0.90, 0 [s = 1-min-max]

Euclidean Distance

Distances = dissimilarities with certain properties

- Euclidean Distance

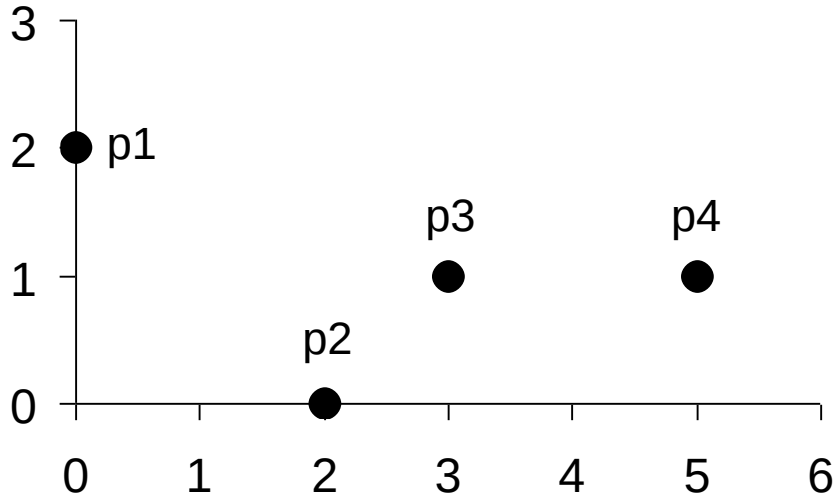
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

- Standardization is necessary, if scales differ.

Euclidean Distance

$d(p1,p2) = SQRT((0-2)^2 + (2-0)^2) = SQRT(4 + 4) = SQRT(8) = 2,828$



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance (Chebyshev).
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

d(p1,p2) = ((|0-2|) + (|2-0|)) = (2 + 2) = 4

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

d(p1,p2) = SQRT($(0-2)^2 + (2-0)^2$) = SQRT(4 + 4) = SQRT(8) = 2,828

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

d(p1,p2) = (max(|0-2|), (|2-0|)) = max(2, 2) = 2

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all x and y and $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

- A distance that satisfies these properties is a metric

Dissimilarity x Distance

- Given two sets A and B, $A - B$ is the set of elements of A that are not in B
  - For example, if $A = \{1, 2, 3, 4\}$ and $B = \{2, 3, 4\}$, then $A - B = \{1\}$ and $B - A = \emptyset$
- We can define the dissimilarity d between two sets A and B as $d(A,B) = size(A - B)$, where $size$ is a function returning the number of elements in a set
  - This dissimilarity, which is an integer value greater than or equal to 0, does not satisfy the second part of the positivity property, the symmetry property, or the triangle inequality

Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$. ($0 \leq s \leq 1$)
   (does not always hold, e.g., cosine)
2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

Similarity Between Binary Vectors

- Common situation is that objects, **x** and **y**, have only binary attributes

- Compute similarities using the following quantities

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

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

  f_{00} = the number of attributes where **x** was 0 and **y** was 0

  f_{11} = the number of attributes where **x** was 1 and **y** was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

  Counts both presences and absences equally – students who had answered questions similarly on a test only of true/false questions

  J = number of 11 matches / number of non-zero attributes

  = $(f_{11}) / (f_{01} + f_{10} + f_{11})$

  Frequently used to handle objects consisting of asymmetric binary attributes – products purchased by customers

SMC versus Jaccard: Example

x = 1 0 0 0 0 0 0 0 0 0

y = 0 0 0 0 0 0 1 0 0 1

f_{01} = 2 (the number of attributes where **x** was 0 and **y** was 1)

f_{10} = 1 (the number of attributes where **x** was 1 and **y** was 0)

f_{00} = 7 (the number of attributes where **x** was 0 and **y** was 0)

f_{11} = 0 (the number of attributes where **x** was 1 and **y** was 1)

SMC = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

= (0+7) / (2+1+0+7) = 0.7

Os alunos responderam 70% da prova da mesma maneira

J = $(f_{11}) / (f_{01} + f_{10} + f_{11})$ = 0 / (2 + 1 + 0) = 0

Os clientes não compraram produtos em comum

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range [0, 1].

2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

$\delta_k = 0$ if the k^{th} attribute is an asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the kth attribute

$\delta_k = 1$ otherwise

3. Compute $$\text{similarity}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use non-negative weights ω_k
  - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$

- Can also define a weighted form of distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

Selecting the Right Proximity Measure

- The type of proximity measure should fit the type of data
  - For many types of dense, continuous data, metric distance measures such as Euclidean distance are often used
  - For sparse data, which often consists of asymmetric attributes, we typically employ similarity measures that ignore 0–0 matches
- In some cases, transformation or normalization of the data is needed to obtain a proper similarity measure
- Practical consideration can also be important
  - Sometimes, one or more proximity measures are already in use in a particular field

Selecting the Right Proximity Measure

- If common practice or practical restrictions do not dictate a choice, then the proper choice of a proximity measure can be a time-consuming task that requires careful consideration of both domain knowledge and the purpose for which the measure is being used
  - A number of different similarity measures may need to be evaluated to see which ones produce results that make the most sense