Home / 1b / 1b.2 / 1b.2.1

1(b).2.1: Measures of Similarity and Dissimilarity

Similarity and Dissimilarity

Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in the literature to compare two data distributions. As the names suggest, a similarity measures how close two distributions are. For multivariate data complex summary methods are developed to answer this question.

Similarity Measure
Numerical measure of how alike two data objects often fall between 0 (no similarity) and 1 (complete similarity)

Dissimilarity Measure
Numerical measure of how different two data objects are range from 0 (objects are alike) to ∞ (objects are different)

Proximity
refers to a similarity or dissimilarity

Similarity/Dissimilarity for Simple Attributes

Here, *p* and *q* are the attribute values for two data objects.

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	$s = 1 - \frac{\lVert p-q \rVert}{n-1}$ (values mapped to integer 0 to n-1, where n is the number of values)	$d = \frac{\lVert p-q \rVert}{n-1}$
Interval or Ratio	$s = 1 - \lVert p-q \rVert, s = \frac{1}{1+\lVert p-q \rVert}$	$d = \lVert p-q \rVert$

Distance, such as the Euclidean distance, is a dissimilarity measure and has some well-known properties: Common Properties of Dissimilarity Measures

1. *d*(*p*, *q*) ≥ 0 for all *p* and *q*, and *d*(*p*, *q*) = 0 if and only if *p* = *q*,
2. *d*(*p*, *q*) = *d*(*q*,*p*) for all *p* and *q*,
3. *d*(*p*, *r*) ≤ *d*(*p*, *q*) + *d*(*q*, *r*) for all *p*, *q*, and r, where *d*(*p*, *q*) is the distance (dissimilarity) between points (data objects), *p* and *q*.

Lessons

Lesson 1(a): Introduction to Data Mining
Lesson 1(b): Exploratory Data Analysis (EDA)
1(b).1 - What is Data
1(b).2 - Numerical Summarization
1(b).2.1: Measures of Similarity and Dissimilarity
1(b).3 - Visualization
1(b).4 - R Scripts
Lesson 2: Statistical Learning and Model Selection
Lesson 3: Linear Regression
Lesson 4: Variable Selection
Lesson 5: Regression Shrinkage Methods
Lesson 6: Principal Components Analysis
Lesson 7: Dimension Reduction Methods
Lesson 8: Modeling Non-linear Relationships
Lesson 9: Classification
Lesson 10: Support Vector Machines
Lesson 11: Tree-based Methods
Lesson 12: Cluster Analysis

Resources

A distance that satisfies these properties is called a **metric**. Following is a list of several common distance measures to compare multivariate data. We will assume that the attributes are all continuous.

Euclidean Distance

Assume that we have measurements x_{ik}, $i = 1, \ldots, N$, on variables $k = 1, \ldots, p$ (also called attributes).

The Euclidean distance between the *i*th and *j*th objects is

$$d_E(i,j) = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

for every pair (i, j) of observations.

The weighted Euclidean distance is:

$$d_{WE}(i,j) = \left(\sum_{k=1}^{p} W_k (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}}$$

If scales of the attributes differ substantially, standardization is necessary.

Minkowski Distance

The Minkowski distance is a generalization of the Euclidean distance.

With the measurement, $x_{ik}, i = 1, \ldots, N, k = 1, \ldots, p$, the Minkowski distance is

$$d_M(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}}$$

where $\lambda \geq 1$. It is also called the L_λ metric.

- $\lambda = 1 : L_1$ metric, Manhattan or City-block distance.
- $\lambda = 2 : L_2$ metric, Euclidean distance.
- $\lambda \rightarrow \infty : L_\infty$ metric, Supremum distance.

$$\lim \lambda \rightarrow \infty = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}} = \max \left(|x_{i1} - x_{j1}|, \ldots, |x_{ip} - x_{jp}| \right)$$

Note that λ and p are two different parameters. Dimension of the data matrix remains finite.

Mahalanobis Distance

Let X be a N × p matrix. Then the i^{th} row of X is

$$x_i^T = (x_{i1}, \ldots, x_{ip})$$

The Mahalanobis distance is

$$d_{MH}(i,j) = \left((x_i - x_j)^T \Sigma^{-1} (x_i - x_j) \right)^{\frac{1}{2}}$$

where $\sum$ is the p×p sample covariance matrix.

Try it!

Calculate the answers to these questions by yourself and then click the icon on the left to reveal the answer.

Analysis of German Credit Data

Analysis of Wine Quality Data

Analysis of Classification Data

Final Project - Sample Work

1. Calculate the Euclidan distances.
2. Calculate the Minkowski distances ($\lambda = 1$ and $\lambda \to \infty$ cases).

1. Calculate the Minkowski distance $\left(\lambda = 1, \lambda = 2, \text{ and } \lambda \to \infty \text{ cases}\right)$ between the first and second objects.
2. Calculate the Mahalanobis distance between the first and second objects.

Common Properties of Similarity Measures

Similarities have some well-known properties:

1. *s*(*p*, *q*) = 1 (or maximum similarity) only if *p* = *q*,
2. *s*(*p*, *q*) = *s*(*q*, *p*) for all *p* and *q*, where *s*(*p*, *q*) is the similarity between data objects, *p* and *q*.

Similarity Between Two Binary Variables

The above similarity or distance measures are appropriate for continuous variables. However, for binary variables a different approach is necessary.

	q=1	q=0
p=1	$n_{1,1}$	$n_{1,0}$
p=0	$n_{0,1}$	$n_{0,0}$

Simple Matching and Jaccard Coefficients

- Simple matching coefficient $= (n_{1,1} + n_{0,0}) / (n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0})$.
- Jaccard coefficient $= n_{1,1} / (n_{1,1} + n_{1,0} + n_{0,1})$.

Try it!

Calculate the answers to the question and then click the icon on the left to reveal the answer.

Given data:

- p = 1 0 0 0 0 0 0 0 0 0
- q = 0 0 0 0 0 0 1 0 0 1

The frequency table is:

	q=1	q=0
p=1	0	1
p=0	2	7

Calculate the Simple matching coefficient and the Jaccard coefficient.

« Previous

Next »