CSC4008 Data Mining

117010279 Ziren WANG

February 2020

0 Getting to Know Your Data

0.1 Data objects & Attribute Types

There are 2 kinds of data, namely categorical data and numerical data:

- **Nominal**: categories, states, or "names of things"
- **Binary**: nominal attribute with only 2 states (0 and 1)
  Symmetric binary: both outcomes equally important. e.g. gender.
  Asymmetric binary: outcomes not equally important. e.g. medical test(positive v.s. negative).
- **Ordinal**: values have a meaningful order (ranking) but magnitude between successive values is not known
- **Numeric**: a measurable quantity, represented in integer or real values
  Interval: measured on a scale of equal-sized units and values have order.
  Ratio: inherent zero-point (an absolute zero). e.g. count.

0.2 Basic Statistical Descriptions of Data

0.2.1 Measuring the Central Tendency

- **mean**: $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **median**: $m = L_1 + (\frac{N/2-(\sum frq)_l}{freq_{median}})width$
- **mode**: the value that occurs most frequently in the set

0.2.2 Measuring the Dispersion of Data

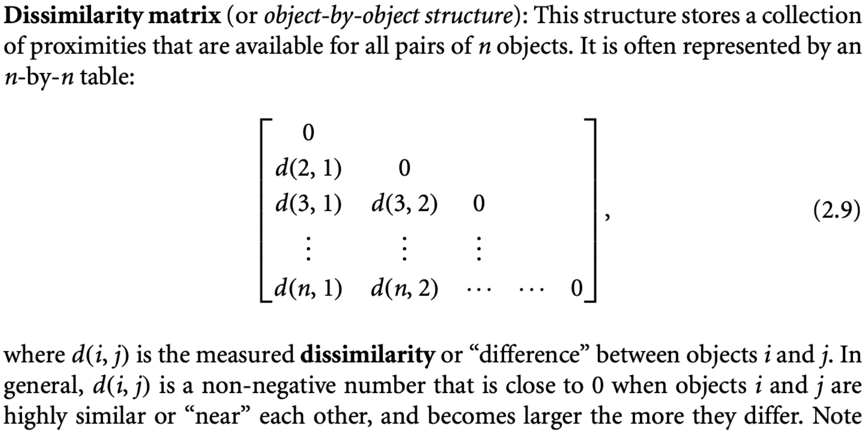
- **range**: max-min
- **quartile**: Q_1, Q_3, points taken at regular intervals of a data distribution, dividing it into essentially equalsize consecutive sets. **Inter-quartile range** : $IQR = Q_3 - Q_1$
- **variance**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = (\frac{1}{N}\sum_{i=1}^{N} x_i^2) - \overline{x}^2$
- **standard deviation**: $\sigma = \sqrt{\sigma^2}$

There is so-called "Five-number Summary" of a distribution: Minimum, Q1, Median, Q3, Maximum.

0.2.3 Graphic Displays of Basic Statistical Descriptions of Data

- **Box Plot**: graphic display of five-number summary;
- **Scatter Plot**: each pair of values is a pair of coordinates and plotted as points in the plane;
- **Quantile Plot**: for a data x_i sorted in increasing order, f_i indicates that approximately f_i100% of the data are below or equal to the value x_i;
- **Quantile-Quantile(QQ) Plot** Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

0.3 Measuring Data Similarity and Dissimilarity



Dissimilarity matrix (or *object-by-object structure*): This structure stores a collection of proximities that are available for all pairs of *n* objects. It is often represented by an *n*-by-*n* table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}, \tag{2.9}$$

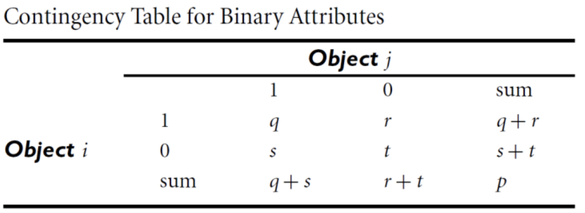
where $d(i, j)$ is the measured **dissimilarity** or "difference" between objects *i* and *j*. In general, $d(i, j)$ is a non-negative number that is close to 0 when objects *i* and *j* are highly similar or "near" each other, and becomes larger the more they differ. Note

Figure 1: Dissimilarity Matrix

0.3.1 Proximity Measures for Nominal Attributes

- **Simple Matching**: $sim(i,j) = 1 - d(i,j) = \frac{m}{p}$, where: m = # of matches and p = total # of variables. Further define $d(i,j) = 1 - sim(i,j) = \frac{p-m}{p}$;
- **Character Coding**: creating a new binary attribute for each of the M nominal states.

0.3.2 Proximity Measures for Binary Attributes



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

Figure 2: Contingency Table for Binary Attributes

Dissimilarity between i and j is: $d(i,j) = \frac{r+s}{q+r+s+t}$. If negative matches t is considered unimportant, then we construct **asymmetric binary dissimilarity** $d(i,j) = \frac{r+s}{q+r+s}$.
Consequently, sim(i,j) of asymmetric binary is called: **Jaccard Coefficient**, calculated by $sim(i,j) = 1 - d(i,j) = \frac{q}{q+r+s}$.

0.3.3 Dissimilarity of Numeric Data

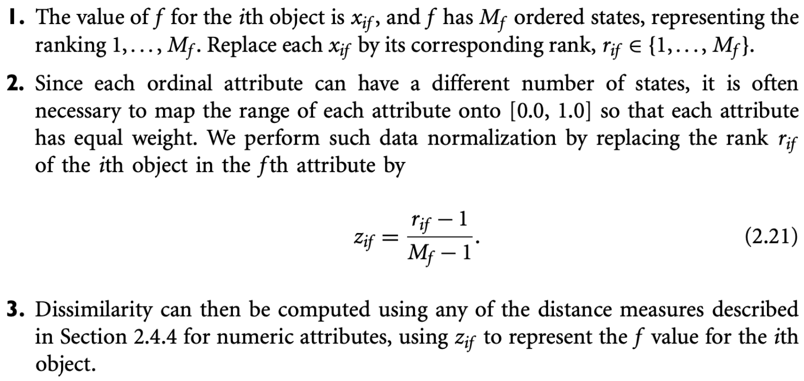
We define Minkowski Distance as: $d(i,j) = (\sum_{f=1}^{p} |x_{if} - x_{jf}|^{h})^{\frac{1}{h}}$. It typically has 4 nice properties:
(1) **Non-negativity**: $d(i,j) \geq 0$, distance is a non-negative number;
(2) **Identity of indiscernibles**: $d(i,i) = 0$, the distance of an object to itself is 0;
(3) **Symmetry**: $d(i,j) = d(j,i)$, distance is a symmetric function;
(4) **Symmetry**: $d(i,j) \leq d(i,k) + d(k,j)$, going directly from object i to object j in space is no more than making a detour over any other object k.
A measure that satisfies these conditions is known as **metric**.

- **Manhattan (city block) distance**: $d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$;
- **Euclidean distance**: $d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$;

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

0.3.4 Proximity Measures for Ordinal Attributes

Suppose that f is an attribute from a set of ordinal attributes describing n objects. The dissimilarity computation with respect to f involves the following steps:



1. The value of f for the ith object is x_{if}, and f has M_f ordered states, representing the ranking $1, \ldots, M_f$. Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots, M_f\}$.
2. Since each ordinal attribute can have a different number of states, it is often necessary to map the range of each attribute onto [0.0, 1.0] so that each attribute has equal weight. We perform such data normalization by replacing the rank r_{if} of the ith object in the fth attribute by

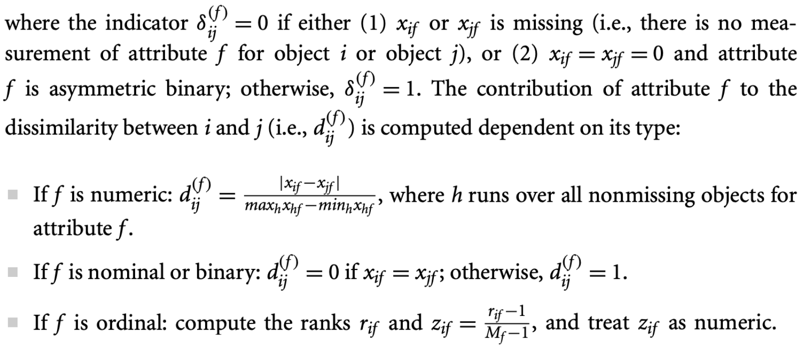
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}. \quad (2.21)$$

3. Dissimilarity can then be computed using any of the distance measures described in Section 2.4.4 for numeric attributes, using z_{if} to represent the f value for the ith object.

Figure 3:

0.3.5 Dissimilarity for Attributes of Mixed Types

Suppose that the data set contains p attributes of mixed type. The dissimilarity $d(i,j)$ between objects i and j is defined as: $d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\delta_{ij}^{(f)}}$.



where the indicator $\delta_{ij}^{(f)} = 0$ if either (1) x_{if} or x_{jf} is missing (i.e., there is no measurement of attribute f for object i or object j), or (2) $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary; otherwise, $\delta_{ij}^{(f)} = 1$. The contribution of attribute f to the dissimilarity between i and j (i.e., $d_{ij}^{(f)}$) is computed dependent on its type:

- If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f.
- If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$.
- If f is ordinal: compute the ranks r_{if} and $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, and treat z_{if} as numeric.

Figure 4:

0.3.6 Cosine Similarity

The traditional distance measures do not work well for the **sparse** numeric data. We need a measure that will focus on the words that the two documents do have in common, meaning, we need a measure for numeric data that ignores zero-matches.
Cosine similarity a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let x and y be two vectors for comparison. Define:

$$sim(x,y) = \frac{x \cdot y}{||x|| |y||}, \text{ where } ||x|| \text{ is the Euclidean norm of vector x.}$$

A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. Note that because the cosine similarity measure does not obey all of the properties defining metric measures, it is referred to as a **nonmetric** measure.