Measuring Similarity

UW ECE 657A - Core Topic

Mark Crowley

Lecture Outline

Measuring Data Similarity and Dissimilarity

Similarity

- A measure of how two data objects are close to each other.
- The more similar the objects the higher the value

Dissimilarity (e.g. distance)

- A measure of how different two data objects are
- The more dissimilar the objects the larger the value of the measure.
- If the two objects are identical then the value is 0

Proximity is used to express similarity or dissimilarity

Why is this important?

Data Matrix and Dissimilarity Matrix

Data (sample) matrix

- n data sample points with d dimensions (features, attributes, ...)

Data Matrix and Dissimilarity Matrix

Data (sample) matrix

- n data sample points with d dimensions (features, attributes, ...)

Distance matrix

- n data points, each is the distance between pairs of points
- A triangular matrix

Data Matrix and Dissimilarity Matrix

Data (sample) matrix

- n data sample points with d dimensions (features, attributes, ...)

$$\begin{bmatrix} x_{11} & \cdots & x_{1r} & \cdots & x_{1d} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ir} & \cdots & xd \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nr} & \cdots & x_{nd} \end{bmatrix}$$

Distance matrix

- n data points, each is the distance between pairs of points
- A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d21 & 0 & & & \\ d31 & d32 & 0 & & \\ : & : & : & & \\ dn1 & dn2 & \cdots & \cdots & 0 \end{bmatrix}$$

9

Measures for Binary Types

Given two binary vectors x and y of d dimensions each.

- **Dot Product:** $x^T y$ measures the number of attributes of value 1 shared between the two vectors
- It can be normalized by the geometric mean of the number of attributes with value one in the vectors $\sqrt{(x^T x)(y^T y)}$
- **Hamming Distance:** measures the number of attributes that have different values in the two vectors (1 in one and 0 in the other and vise versa).
- Equilvalent to $\sum_{i=1}^{d} |x_i - y_i|$

Measures for Binary Types

- **Tanimoto Measure:** it measures the common attributes relative to the non-common elements

$$t(x,y) = \frac{x^T y}{x^T x + y^T y - x^T y}$$

-

Contingency Tables

Contingency tables can be used to define of number of other proximity measures.

Contingency table:

		y	
		1	0
x	1	m11	m10
	0	m01	m00

Distance Coefficients

Simple Matching Coefficient (SMC):

$$\frac{m_{00} + m_{11}}{m_{00} + m_{11} + m_{01} + m_{10}}$$

Similarity with equal weight given to 0 and 1.
Jaccard Coefficient (JC):

$$\frac{m_{11}}{m_{11} + m_{01} + m_{10}}$$

Similarity ignoring 0-0 matches

- SMC and JC always less than 1
- value of 1 means the two vectors are identical
- (1 − SMC) yields distance
- (1 − JC) yields Jaccard distance

Measures of Nominal Types

- **Converting Nominal Type to Binary Type:** for each of the nominal values of the variables define as a binary variable. Binary variable is 1 if variable has corresponding value, 0. Then we can use the binary measures.
- Can use the number of matches of the values of attributes relative to the number of attributes as simple similarity measure and 1-similarty as distance.

Definition: Distance Metric

- **Distance metric:** For all vectors x, y and z the function d is a metric *iff*:
  - $d(x, x) = 0$ where $d(x, y) = 0$ iff $x = y$
  - $d(x, y) \geq 0$ Non-negativity
  - $d(x, y) = d(y, x)$ Symmetry
  - $d(x, y) \leq d(x, z) + d(z, y)$ Triangle inequality
- e.g. **Minkowski metric**:

$$d_k(x, y) = \left[\sum_{i=1}^{n} |x_i - y_i|^k \right]^{1/k}$$

Minkowksi Metric

$$d_k(x, y) = \left[\sum_{i=1}^{n} |x_i - y_i|^k \right]^{1/k}$$

- For k=2, we get the l_2 norm or Euclidean distance
- For k=1, we get the l_1 norm. Also called absolute norm, or city block norm or *Manhattan distance*.
- For $k \to \infty$, we get the supremum or Chebyshev distance

$$d_\infty(x, y) = \max_i |x_i - y_i|$$

- $k \to -\infty$ we get the min distance

$$\min_i |x_i - y_i|$$

Mahalanobis Distance

Question:

After you compute the distance between two points, how do you know if that distance is *significant* or *relevant*?

Mahalanobis Distance

Question:

After you compute the distance between two points, how do you know if that distance is *significant* or *relevant*?

Mahalanobis Distance is another way to measure difference between vectors that accounts for their *covariance*:

$$d(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Where x and y share the *same* distribution and covariance matrix S.

Mahalanobis Distance

Question:

After you compute the distance between two points, how do you know if that distance is *significant* or *relevant*?

Mahalanobis Distance is another way to measure difference between vectors that accounts for their *covariance*:

$$d(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Where x and y share the *same* distribution and covariance matrix S.

Interpretation:

- Multi-dimensional generalization of distance
- Accounting for how many standard deviations away X is from the mean of Y.
- Distance takes into account the disimilarity between two vectors given the overall dataset covariance.

Mahalanobis Distance : Properties

Mahalanobis Distance: $d(x, y) = \sqrt{(x - y)^T S^{-1}(x - y)}$

Properties:

- Distance is preserved under linear transformations of data.
- If $S = I$, then all datapoints are perfectly correlated, and the Mahalanobis distance is equivalent to the Euclidean distance.
- In other words, *Euclidean distances are normalized to essentially be in "units" of standard deviation for the whole dataset*.

Mahalanobis Distance : Properties

Mahalanobis Distance: $d(x, y) = \sqrt{(x-y)^T S^{-1}(x-y)}$

If the covariance matrix is diagonal, then we obtain the **standardized Euclidean distance**:

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

where s_i is the standard deviation of the x_i and y_i over the sample set.

Measures for Ordinal Types

- Ordinal values are ranked values
- Can be mapped to the interval type
  - replace x_{if} by their rank $r_{if} \in \{1, ..., M_f\}$
  - map the range of each variable onto $[0, 1]$ by replacing i^{th} object in the f^{th} variable by

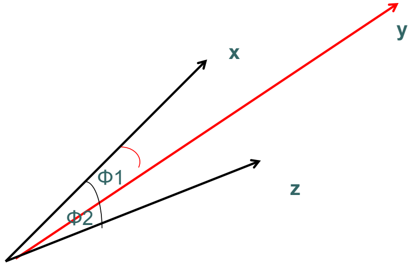
$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - One can then use methods for interval-scaled variables

Cosine Similarity

- The similarity between two vectors x and y is measured as the cosine of the angle of the two vectors: $\cos(x, y) = x^T y / ||x|| ||y||$
- i.e. the dot product of the two vectors normalized by the product of their lengths.
- Value range from -1 (opposite) to 1 same(except for length).
- It is a useful measure for similarity that is widely used in information retrieval and data mining especially for sparse vectors.
- Note that the measure focuses on the shared non-zero attribute values and ignores any 0-* matches between the two vectors.
- If the vectors are binary it reduces to the number of attributes (with value 1) shared by the two vectors normalized by the geometric mean of their sizes.

Cosine Similarity



x is closer to y than z using the cosine similarity measure

Example

$$\text{Let } x = (0,0,1,1,3,0,1)$$
$$y = (0,4,3,1,1,0,0)$$
$$z = (1,3,1,0,0,1,0)$$

$$||x|| = \sqrt{1+1+9+1} = 3.5$$
$$||y|| = \sqrt{16+9+1+1} = 5.2$$
$$||z|| = \sqrt{1+9+1+1} = 3.5$$

$$x^T y = 3+1+3 = 7$$
$$x^T z = 1$$

$$\cos(x,y) = 7/18.2 = 0.38$$
$$\cos(x,z) = 1/12.25 = 0.08$$

$$||x-y||_2 = 5$$
$$||x-z||_2 = 4.6$$