Introduction to Machine Learning

Ranga Raju Vatsavai, Ph.D.
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)

Feb. 25-27, 2019

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
- Proximity refers to a similarity or dissimilarity

Attributes

- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Recall From EDA

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit. (location of zero is arbitrary)
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts (location of zero is fixed)

Similarity/Dissimilarity for Simple Attributes

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

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

Table 5.1. Similarity and dissimilarity for simple attributes

Normalization Vs. Standardization

- Rescaling
  - + or - by Constant and * or / by constant
- Normalization
  - Rescales the values into [0,1]

$$X_o = \frac{(X_i - X_{\min})}{(X_{\max} - X_{\min})}$$

- Standardization
  - Rescales data to have 0 mean and 1 sd

$$X_o = \frac{(X_i - \mu)}{(\sigma)}$$

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation s_x is the most common measure of the spread of a set of points

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- Because of outliers, othe measures ar often used

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}\lvert x_i - \overline{x}\rvert$$

$$\text{MAD}(x) = median\Big(\{\lvert x_1 - \overline{x}\rvert, \ldots, \lvert x_m - \overline{x}\rvert\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Euclidean Distance

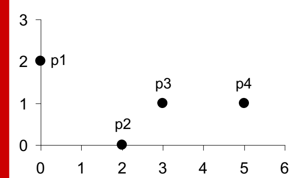
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.

Euclidean Distance



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

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) of data objects p and q.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this 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.
  - 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

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

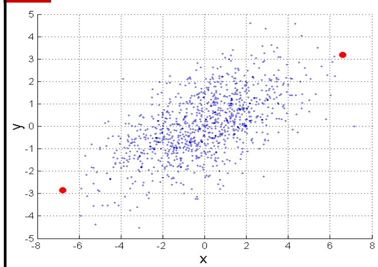
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Mahalanobis Distance

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$



Σ is the covariance matrix of the input data X

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

Common Properties of a Distance

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

  1. *d(p, q)* ≥ 0 for all *p* and *q* and *d(p, q) = 0* only if *p = q*. (Positive definiteness)
  2. *d(p, q) = d(q, p)* for all *p* and *q*. (Symmetry)
  3. d(*p, r*) $\leq$ *d(p, q) + d(q, r)* for all points *p, q,* and *r.* (Triangle Inequality)

  where *d(p, q)* is the distance (dissimilarity) between points (data objects), *p* and *q*.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

- Similarities, also 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*. (Symmetry)

  where *s(p, q)* is the similarity between points (data objects), *p* and *q*.

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities
  - F01 = the number of attributes where *p* was 0 and *q* was 1
  - F10 = the number of attributes where *p* was 1 and *q* was 0
  - F00 = the number of attributes where *p* was 0 and *q* was 0
  - F11 = the number of attributes where *p* was 1 and *q* was 1
- Simple Matching and Jaccard Coefficients
  - SMC = number of matches / number of attributes
    = (F11 + F00) / (F01 + F10 + F11 + F00)
  - J = number of 11 matches / number of non-zero attributes
    = (F11) / (F01 + F10 + F11)

SMC versus Jaccard: Example

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

F01 = 2 (the number of attributes where *p* was 0 and *q* was 1)
F10 = 1 (the number of attributes where *p* was 1 and *q* was 0)
F00 = 7 (the number of attributes where *p* was 0 and *q* was 0)
F11 = 0 (the number of attributes where *p* was 1 and *q* was 1)

SMC = (F11 + F00) / (F01 + F10 + F11 + F00)
= (0+7) / (2+1+0+7) = 0.7

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

Cosine Similarity

- If d_1 and d_2 are two document vectors, then
  cos(*A*, *B*) = (*A* • *B*) / ||*A*|| ||*B*|| ,
  where • indicates vector dot product and ||*A*|| is the length of vector *A*.

$$A \cdot B = \sum_{i=1}^{n} A_i B_i = A_1 B_1 + A_2 B_2 + \ldots + A_n B_n$$

$$\| A \| = \sqrt{A \cdot A}$$

- Example:
  ***A* = 3 2 0 5 0 0 0 2 0 0**
  ***B* = 1 0 0 0 0 0 0 1 0 2**

A • *B*= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
||*A*|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481
||*B*|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.449
cos(*A*, *B*) = .3150

Extended Jaccard Coefficient (Tanimoto)

- Variation of Jaccard for continuous or count attributes

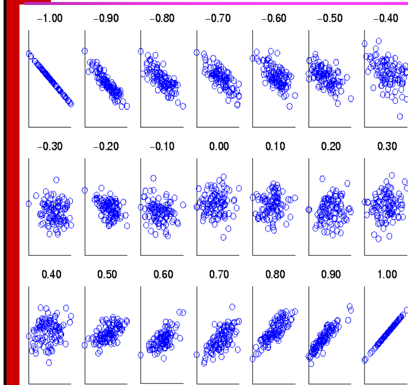
$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$
$$q'_k = (q_k - mean(q)) / std(q)$$
$$correlation(p, q) = p' \bullet q' / (n-1)$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Drawback of Correlation

- X = (-3, -2, -1, 0, 1, 2, 3)
- Y = (9, 4, 1, 0, 1, 4, 9) $Y = X^2$

- Mean(X) = 0, Mean(Y) = 4
- Std(X) = 2.16, Std(Y) = 3.74

- Correlation
  = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (2.16 * 3.74)
  = 0