Similarity and Dissimilarity Measures

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

distance

- 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

1. objects having only one simple attribute
2. objects with multiple attributes

Similarity/Dissimilarity for Simple Attributes

The following table shows the similarity and dissimilarity between two objects, *x* and *y*, with respect to a single 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	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Dissimilarities between Data Objects

various kinds of dissimilarities

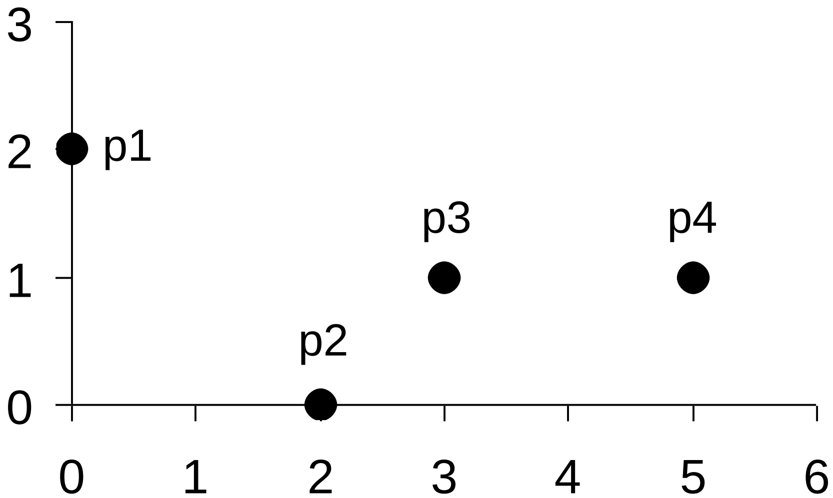
1. distances, which are dissimilarities with certain properties
2. provide examples of more general kinds of dissimilarities

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 $\mathbf{y}$.

Distances:example



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

Distances

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 $\boldsymbol{y}$.

Distances

Minkowski Distance

- $r = 1$. 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

Distances:Example

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

Standardization and Correlation for Distance Measures

A generalization of Euclidean distance, the **Mahalanobis distance** when attributes are correlated, have different ranges of values

$$\mathbf{mahalanobis}(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})$$

Σ is the covariance matrix

ij_{th} entry is the covariance of the i_{th} and j_{th} attributes

Distances

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

1. $d(x, y) \geq 0$ for all x and y
2. $d(x, y) = 0$ only if $x = y$
3. $d(x, y) = d(y, x)$ for all x and y.
4. $d(x, z) \leq d(x, y) + d(y, z)$ for all points x, y, and z.

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

Metric

Non-metric Dissimilarities

some dissimilarities do not satisfy one or more of the metric properties

Example: Non-metric Dissimilarities: Set Differences

Given two sets *A* and *B*, *A* − *B* is the set of elements of A that are not in *B*.

$A = \{1, 2, 3, 4\}$ and $B = \{2, 3, 4\}$ $\quad A - B = \{1\}$ $\quad B - A = \emptyset$

$$d(A, B) = size(A - B)$$

$$d(A, B) = size(A - B) + size(B - A)$$

Similarities between Data Objects

Similarities, have some typical properties.

1. $s(x, y) = 1$ (or maximum similarity) only if $x = y$.

2. $s(x, y) = s(y, x)$ for all x and y. (Symmetry)

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

Similarities between Data Objects

Similarity Between Binary Vectors

p and q, have only binary attributes

Compute similarities using the following quantities

f_{01} = the number of attributes where p was 0 and q was 1

f_{10} = the number of attributes where p was 1 and q was 0

f_{00} = the number of attributes where p was 0 and q was 0

f_{11} = the number of attributes where p was 1 and q was 1

Simple Matching Coefficient

SMC = number of matches / number of attributes

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

Jaccard Coefficient

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

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

Similarities between Data Objects

Example: binary similarity

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 p was 0 and q was 1)
$f_{10} = 1$ (the number of attributes where p was 1 and q was 0)
$f_{00} = 7$ (the number of attributes where p was 0 and q was 0)
$f_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$\text{SMC} = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

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

Similarities between Data Objects

Cosine Similarity

If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d_1}, \mathbf{d_2}) = \langle \mathbf{d_1},\mathbf{d_2} \rangle / \|\mathbf{d}_1\| \, \|\mathbf{d}_2\| ,$$

Example:

$$\mathbf{d_1} = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$\mathbf{d_2} = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$\langle \mathbf{d_1}, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$| \mathbf{d_1} \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}} = 6.481$

$\| \mathbf{d_2} \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{\mathbf{0.5}} = (6)^{\mathbf{0.5}} = 2.449$

$\cos(\mathbf{d_1}, \mathbf{d_2}) = 0.3150$

Similarities between Data Objects

Correlation

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}$$

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

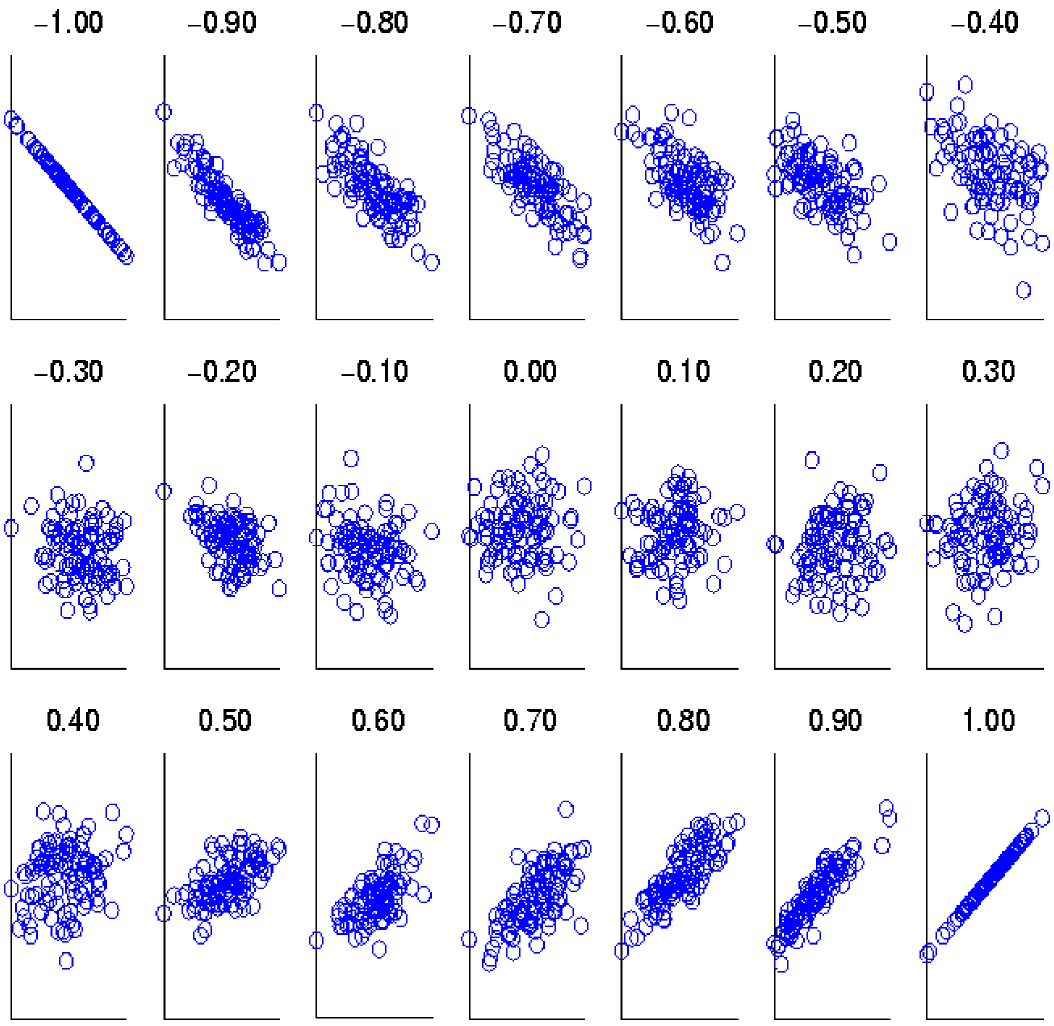
$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

$$s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

Similarities between Data Objects



Scatter plots showing the similarity from –1 to 1.

Similarities between Data Objects

Drawback of Correlation : **Non-linear Relationships**

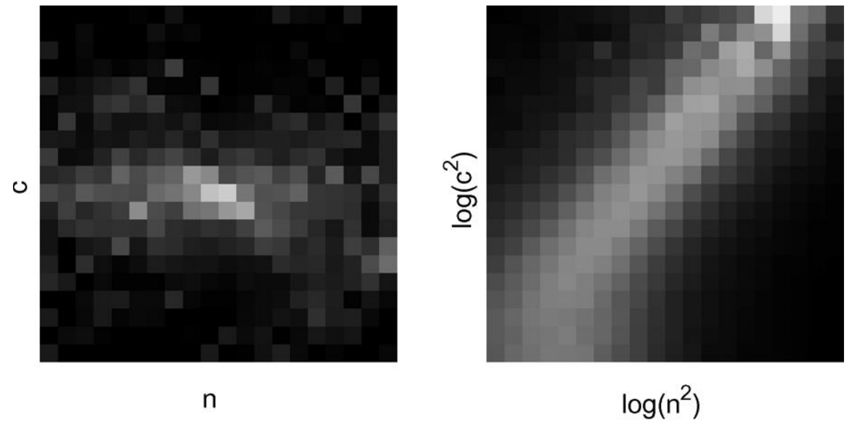
- correlation is 0, no linear relationship between the attributes of the two data objects
- non-linear relationships may still exist.

Example:

$\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
$\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$
$y_i = x_i^2$
$\text{mean}(\mathbf{x}) = 0$, $\text{mean}(\mathbf{y}) = 4$
$\text{std}(\mathbf{x}) = 2.16$, $\text{std}(\mathbf{y}) = 3.74$

$\text{corr} = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) \;/\; (\,6 * 2.16 * 3.74\,) = 0$

Similarities between Data Objects



(a) (b)

Information theory

- Information theory
- similarity measures
- handle non-linear relationships
- complicated and time intensive to compute
- Information relates to possible outcomes of an event
- information is related the probability of an outcome
- The smaller the probability of an outcome, the more information it provides
- Entropy is the commonly used measure

Entropy

- ✓ a variable (event), X,
- ✓ with n possible values (outcomes), $x_1, x_2 \ldots, x_n$
- ✓ each outcome having probability, $p_1, p_2 \ldots, p_n$
- ✓ the entropy of X , $H(X)$, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Entropy is between 0 and $\log_2 n$ and is measured in bits

entropy is a measure of how many bits it takes to represent an observation of X on average

Example:

For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

For $p = 0.5$, $q = 0.5$ (fair coin) $H = 1$

For $p = 1$ or $q = 1$, $H = 0$

Entropy

a number of observations (m) of some attribute, X, e.g., the hair color of students in the class, where there are n different possible values

the number of observation in the i^{th} category is m_i

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

Hair Color	Count
Black	75
Brown	15
Blond	5
Red	0
Other	5
Total	100

Maximum entropy is $\log_2 5 = 2.3219$

Mutual Information

Information one variable provides about another

Formally, $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where

$H(X,Y)$ is the joint entropy of X and Y,

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

✓how similar the joint distribution *p(X, Y) is to the factored distribution p(X)p(Y).*

MI is zero iff the variables are independent

MI between *X and Y as the reduction in uncertainty about X after* observing *Y*

Mutual Information example

Student Status	Count
Undergrad	45
Grad	55
Total	100

Grade	Count
A	35
B	50
C	15
Total	100

Student Status	Grade	Count
Undergrad	A	5
Undergrad	B	30
Undergrad	C	10
Grad	A	30
Grad	B	20
Grad	C	5
Total		100

Mutual information of Student Status and Grade = 0.9928 + 1.4406 - 2.2710 = 0.1624