Master Degree in Artificial Intelligence for Science and Technology

Proximity Measures

Fabio Stella
Department of Informatics, Systems and Communications
University of Milano-Bicocca
fabio.stella@unimib.it

OUTLOOK

- SIMILARITY – DISSIMILARITY
- EUCLIDEAN – MINKOWSKI – CITY BLOCK – MAHALANOBIS
- SIMPLE MATCHING COEFFICIENT – JACCARD – COSINE SIMILARITY
- CORRELATION
- ENTROPY – MUTUAL INFORMATION

- **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

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

EUCLIDEAN DISTANCE

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

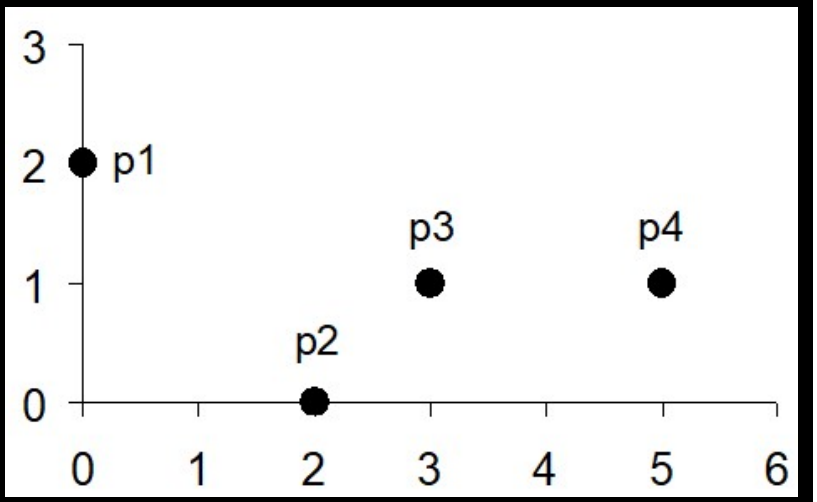
where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) of data objects **x** and **y**.

— Standardization is necessary, if scales differ

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Distance between 2 points x and y

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$



	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

Distance matrix is symmetric

MINKOWSKI DISTANCE is a generalization of euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

for r=2 -> euclidean

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) of data objects **x** and **y**.

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

- r = 1. **CITY BLOCK** (**MANHATTAN**, **TAXICAB**, L_1 norm) distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k| \right)$$

A common example of this for binary vectors is the **HAMMING DISTANCE**, which is just the number of bits that are different between two binary vectors

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

$$(0, 0, 1, 0, 0, 1, 0, 1, 1, 0)$$
$$(1, 0, 1, 1, 0, 0, 0, 0, 1, 0)$$

$$1 + 0 + 0 + 1 + 0 + 1 + 0 + 1 + 0 + 1 = 5$$

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r\right)^{1/r}$$

- $r = 2$. **EUCLIDEAN DISTANCE**

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

- $r \to \infty$. "**SUPREMUM**" (L_{max} norm, L_∞ norm) distance

This is the maximum difference between any component of the vectors

$$d(\mathbf{x},\mathbf{y}) = \underset{k}{\operatorname{argmax}} |x_k - y_k|$$

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

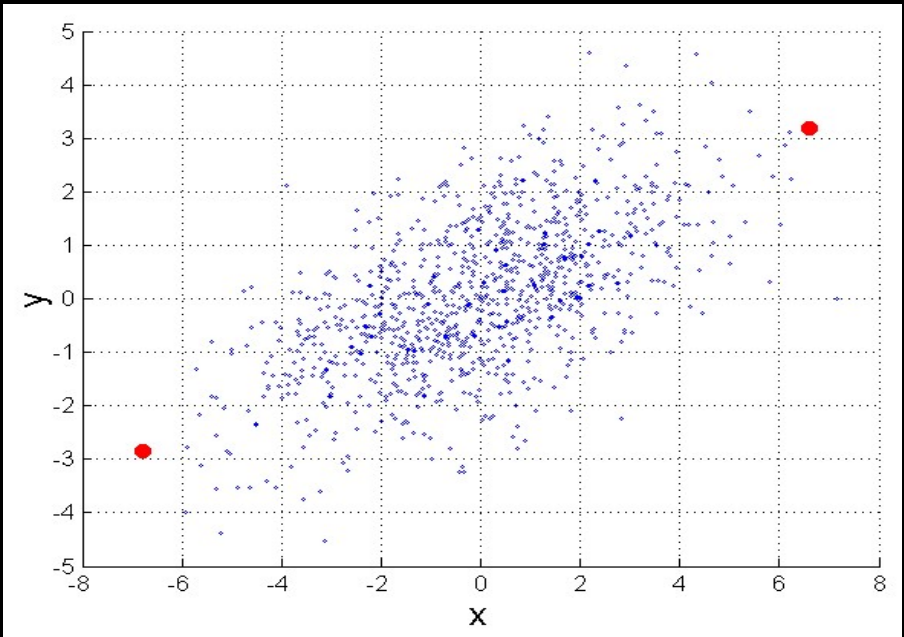
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

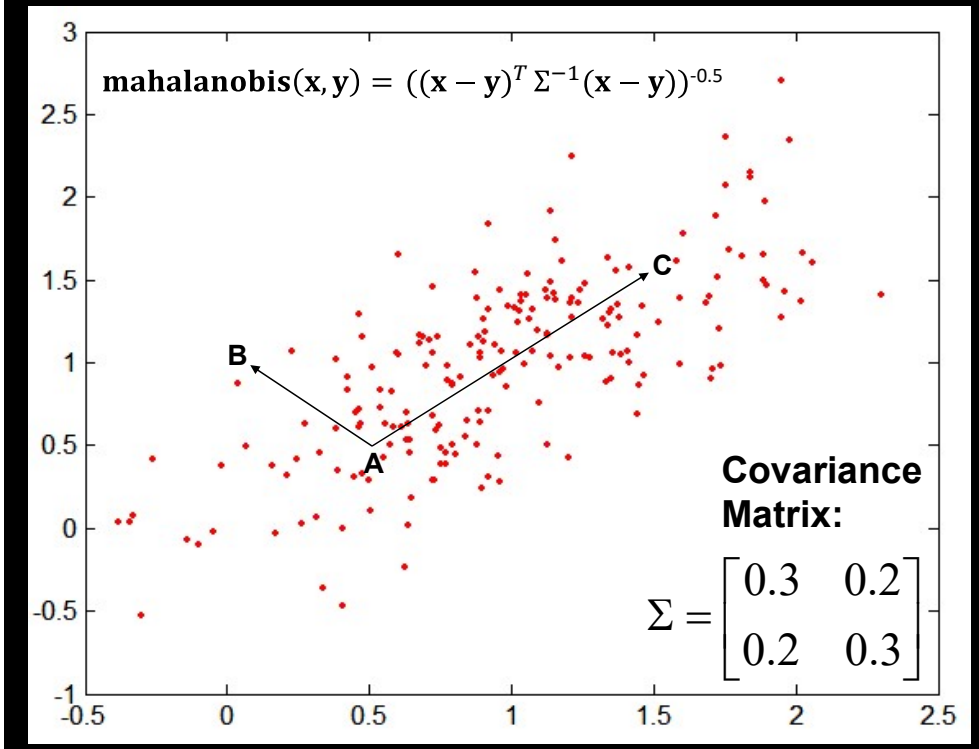
MAHALANOBIS DISTANCE

$$\mathbf{mahalanobis}(\mathbf{x},\mathbf{y}) = ((\mathbf{x}-\mathbf{y})^T \Sigma^{-1}(\mathbf{x}-\mathbf{y}))^{-0.5}$$



Σ is the covariance matrix

For red points, the euclidean distance is 14.7, Mahalanobis distance is 6.



A = (0.5, 0.5)

B = (0, 1)

C = (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

DISTANCES, such as the Euclidean distance, have some well known **PROPERTIES**.

— $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}$ and $\mathbf{y}$ and $d(\mathbf{x}, \mathbf{y}) = 0$ iff $\mathbf{x} = \mathbf{y}$ **NON NEGATIVE**

— $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$ **SYMMETRY**

— $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ **TRIANGLE INEQUALITY**

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$

A distance that satisfies these properties is a **METRIC**

SIMILARITIES, also have some well known **PROPERTIES**.

— $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$ (does not always hold, e.g., cosine)

— $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$ **SYMMETRY**

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

- Common situation is that objects, **x** and **y**, have only binary attributes
- Compute similarities using the following quantities

  f_{01} = the number of attributes where **x** was 0 and **y** was 1

  f_{10} = the number of attributes where **x** was 1 and **y** was 0

  f_{00} = the number of attributes where **x** was 0 and **y** was 0

  f_{11} = the number of attributes where **x** was 1 and **y** was 1

- **SIMPLE MATCHING COEFFICIENT**

  SMC = number of matches / number of attributes = $\frac{f_{11}+f_{00}}{f_{01}+f_{10}+f_{11}+f_{00}}$

- **JACCARD COEFFICIENT**

  J = number of 11 matches / number of non-zero attributes = $\frac{f_{11}}{f_{01}+f_{10}+f_{11}}$

$$\mathbf{x} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\mathbf{y} = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$$

f_{01} = 2 (the number of attributes where **x** was 0 and **y** was 1)

f_{10} = 1 (the number of attributes where **x** was 1 and **y** was 0)

f_{00} = 7 (the number of attributes where **x** was 0 and **y** was 0)

f_{11} = 0 (the number of attributes where **x** was 1 and **y** was 1)

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

$$\text{J} = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+1+0} = 0$$

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

$$\cos(\mathbf{d}_1,\mathbf{d}_2) = \langle \mathbf{d}_1,\mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \, \|\mathbf{d}_2\|,$$ **COSINE SIMILARITY**

where $\langle \mathbf{d}_1,\mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\|\mathbf{d}\|$ is the length of vector $\mathbf{d}$.

EXAMPLE: $\mathbf{d}_1 = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$ $\mathbf{d}_2 = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$

$\langle \mathbf{d}_1,\mathbf{d2} \rangle = 3 \bullet 1 + 2 \bullet 0 + 0 \bullet 0 + 5 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 2 \bullet 1 + 0 \bullet 0 + 0 \bullet 2 = 5$

$\| \mathbf{d}_1 \| = (3 \bullet 3 + 2 \bullet 2 + 0 \bullet 0 + 5 \bullet 5 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 2 \bullet 2 + 0 \bullet 0 + 0 \bullet 0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d}_2 \| = (1 \bullet 1 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 0 \bullet 0 + 1 \bullet 1 + 0 \bullet 0 + 2 \bullet 2)^{0.5} = (6)^{0.5} = 2.449$

$\cos(\mathbf{d}_1,\mathbf{d}_2) = \langle \mathbf{d}_1,\mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \, \|\mathbf{d}_2\| = 5 / (6.481 \bullet 2.449) = 0.3150$

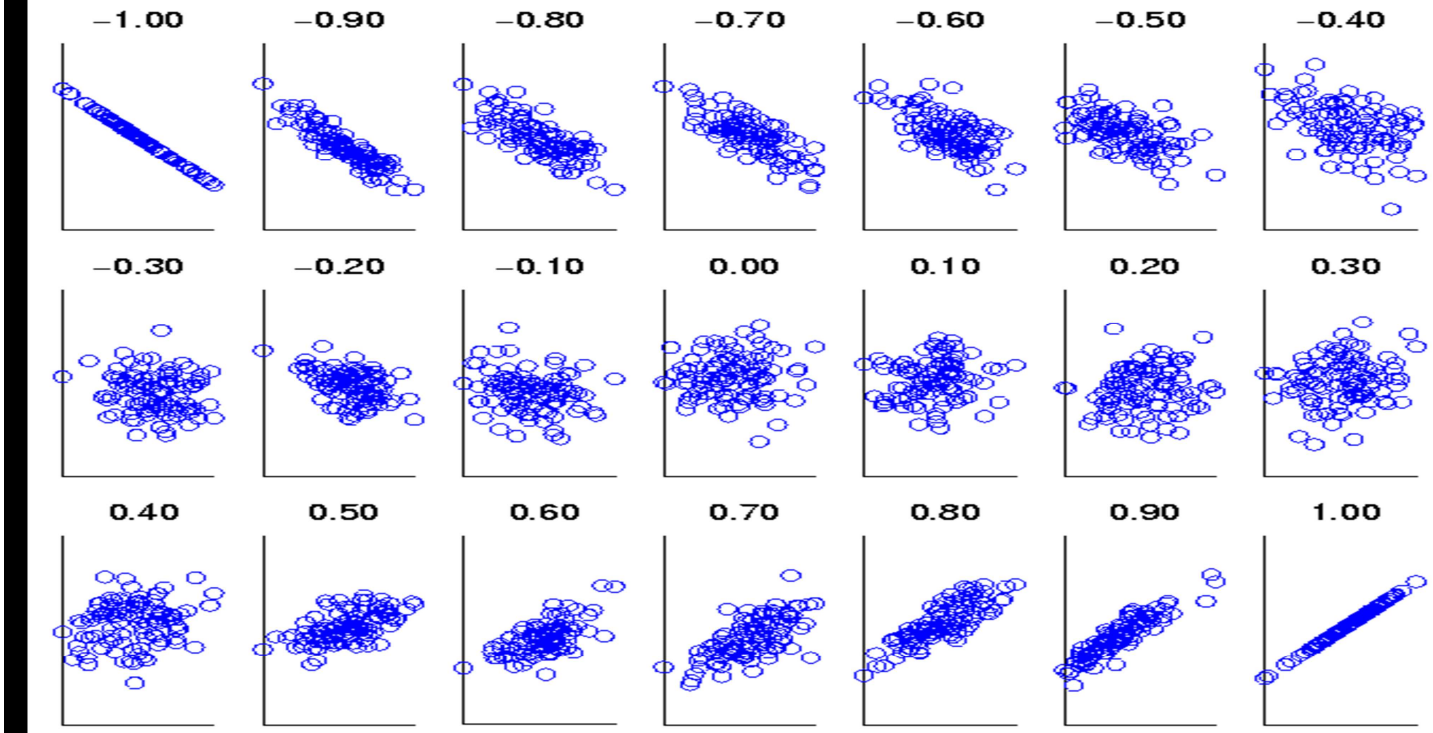
CORRELATION

$$Corr(\mathbf{x}, \mathbf{y}) = \frac{Cov(\mathbf{x}, \mathbf{y})}{\sqrt{Var(\mathbf{x})Var(\mathbf{y})}} = \frac{s_{\mathbf{xy}}}{s_{\mathbf{x}}\, s_{\mathbf{y}}}$$

$\mathbf{x}, \mathbf{y}$, n-dimensional vectors

$$s_{\mathbf{xy}} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})\,(\mathbf{y}_k - \bar{\mathbf{y}}) \qquad s_{\mathbf{x}} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})^2} \qquad s_{\mathbf{y}} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{y}_k - \bar{\mathbf{y}})^2}$$

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k \qquad \bar{\mathbf{y}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{y}_k$$


−1.00
−0.90
−0.80
−0.70
−0.60
−0.50
−0.40
−0.30
−0.20
−0.10
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

DRAWBACK OF CORRELATION

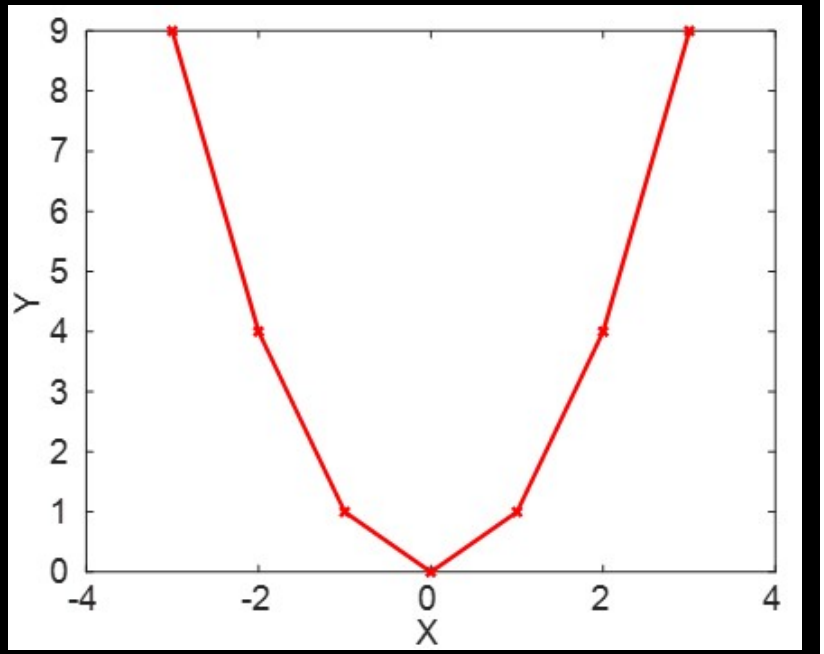
x = (-3, -2, -1, 0, 1, 2, 3)

y = (9, 4, 1, 0, 1, 4, 9)

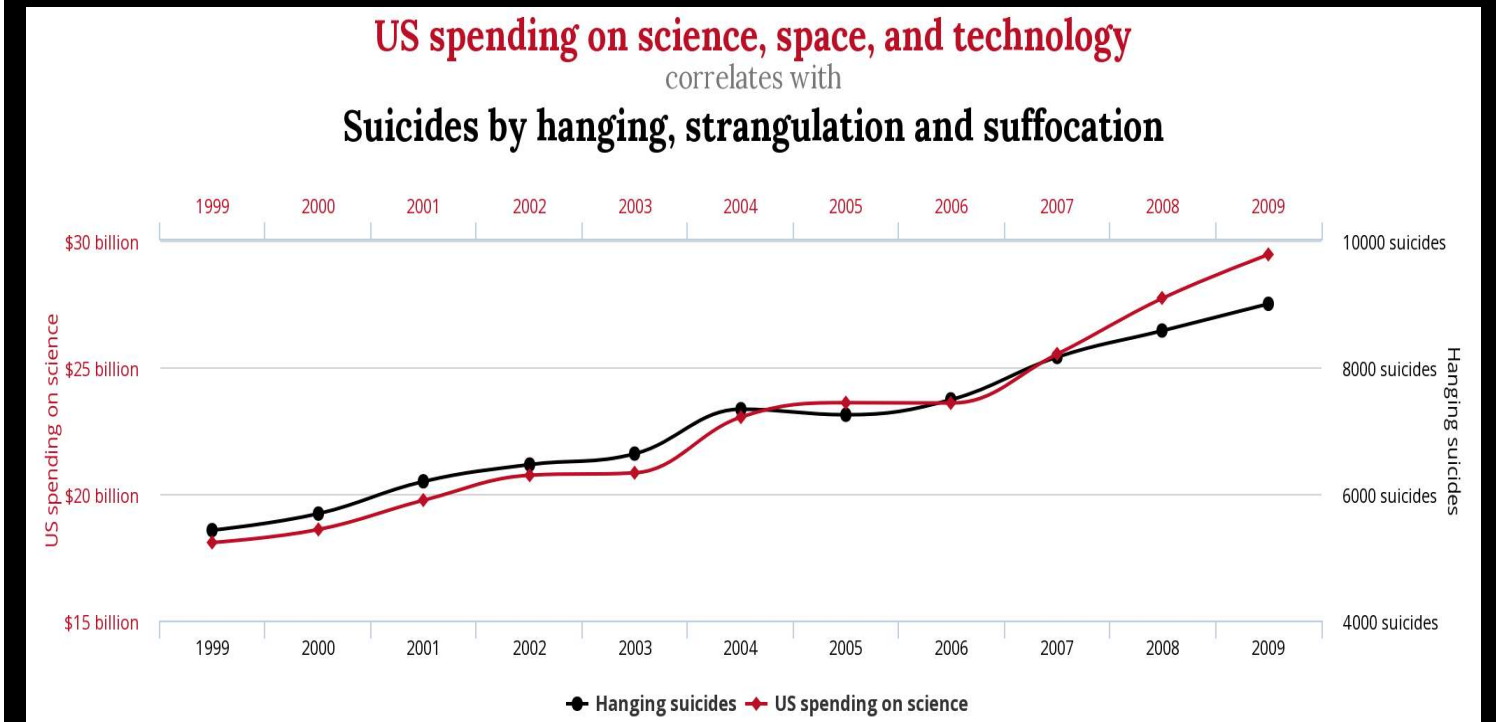
$y_i = x_i^2$

$\bar{\mathbf{x}} = 0, \bar{\mathbf{y}} = 4$

$s_{\mathbf{x}} = 2.16, s_{\mathbf{y}} = 3.74$



$Corr(\mathbf{x}, \mathbf{y})$ = (-3)•(5) + (-2)•(0) + (-1)•(-3) + (0)•(-4) + (1)•(-3) + (2)•(0) + (3)•(5) / (6 • 2.16 • 3.74) = 0


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Hanging suicides
US spending on science
tylervigen.com

Compare the three proximity measures according to their behavior under variable transformation

- **SCALING**: multiplication by a value
- **TRANSLATION**: adding a constant

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

EXAMPLE:

- $\mathbf{x}$ = (1, 2, 4, 3, 0, 0, 0),
- $\mathbf{y}$ = (1, 2, 3, 4, 0, 0, 0)
- $\mathbf{y_s} = \mathbf{y} \cdot \mathbf{2}$ (scaled version of $\mathbf{y}$),
- $\mathbf{y_t} = \mathbf{y} + \mathbf{5}$ (translated version of $\mathbf{y}$)

Measure	(x, y)	(x, y_s)	(x, y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Choice of the right proximity measure depends on the domain

- What is the correct choice of proximity measure for the following situations?
  - — Comparing documents using the frequencies of words
    - documents are considered similar if the word frequencies are similar
  - — Comparing the temperature in Celsius of two locations
    - two locations are considered similar if the temperatures are similar in magnitude
  - — Comparing two time series of temperature measured in Celsius
    - two time series are considered similar if their “shape” is similar, i.e., they vary in the same way over time, achieving minimums and maximums at similar times, etc…

- Domain of Application
  - similarity measures tend to be specific to the type of attribute and data
  - record data, images, graphs, sequences, 3D-protein structure, etc. tend to have different measures
- However, one can talk about various properties that you would like a proximity measure to have:
  - symmetry is a common one
  - tolerance to noise and outliers is another
  - ability to find more types of patterns?
  - many others possible
- The measure must be applicable to the data and produce results that agree with domain knowledge

- **INFORMATION THEORY** is a well-developed and fundamental discipline with broad applications.
- Some similarity measures are based on information theory:
  - **MUTUAL INFORMATION** in various versions
  - **MAXIMAL INFORMATION COEFFICIENT** (MIC) and related measures
  - general and can handle non-linear relationships
  - can be complicated and time intensive to compute

- Information relates to possible outcomes of an event
  - transmission of a message, flip of a coin, or measurement of a piece of data
- The more certain an outcome, the less information that it contains and vice-versa
  - For example, if a coin has two heads, then an outcome of heads provides no information
  - More quantitatively, the information is related the probability of an outcome
    - The smaller the probability of an outcome, the more information it provides and vice-versa
  - Entropy is the commonly used measure

ENTROPY, for

- a variable (event), X,
- with *n* possible values (outcomes), x_1, x_2 …, x_n
- each outcome having probability, p_1, p_2 …, 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
  - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Given a coin $X \in \{head, tail\}$

$p = probability\ of\ X = head$

$1 - p = probability\ of\ X = tail$

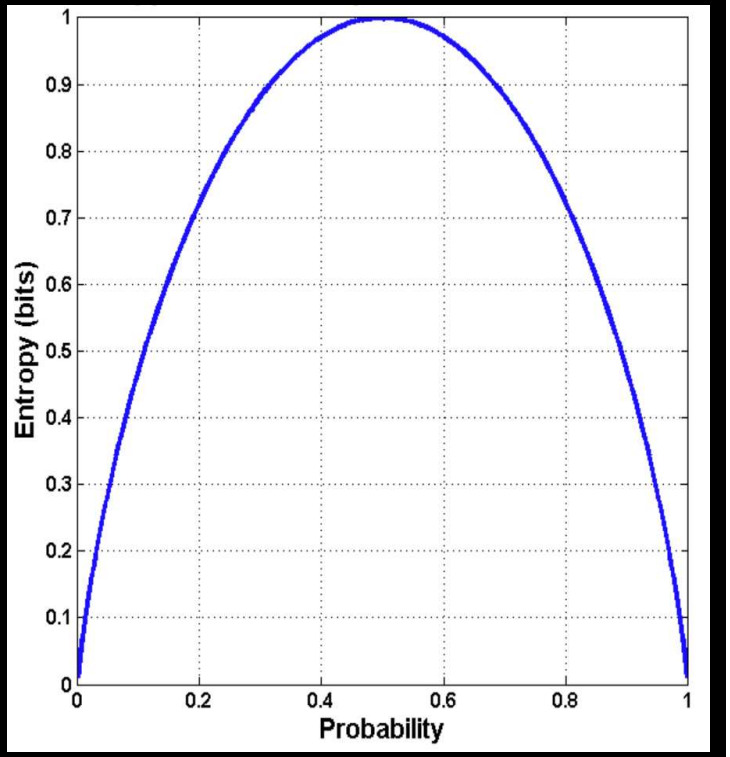
$$H(X) = -\sum_{j=1}^{m} P(X = x_j)\ log_2 P(X = x_j)$$

$$H(X) = -\sum_{j=1}^{2} P(X = x_j)\ log_2 P(X = x_j)$$

$$= -p\ log_2 p - (1 - p)\ log_2 (1 - p)$$

$p = 0.5$ (fair coin) $\Longrightarrow H(X) = 1$ $(max\ uncertainty)$

$p = 1.0$ $\Longrightarrow H(X) = 0$ $(min\ uncertainty)$



Suppose we have

- 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
- and the number of observation in the i^{th} category is m_i
- then, for this sample

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

Hair Color	Count	p	$-p\log_2 p$
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	100	1.0	1.1540

$$Hair\ Color\ \ X \in \{Black, Brown, Blond, Red, Other\}$$

$$H(X) = -\sum_{j=1}^{5} P(X = x_j)\ log_2 P(X = x_j) = 1.154$$

$$\max H(X) = log_2 5 = 2.3219$$

For continuous attributes the computation is more difficult, i.e., we could apply discretization.

MUTUAL INFORMATION is used to measure the similarity between two sets of paired values.

- sometimes used in place of correlation, particularly when a nonlinear relationship is suspected between the pair of values
- a measure of how much information one set of values provides about another, given that the values come in pairs, e.g., height and weight
- if the two sets of values are independent, i.e., the value of one tells us nothing about the other, then their mutual information is equal to 0. On the other hand, if the two sets of values are completely dependent, i.e., knowing the value of one tells us the value of the other and vice-versa, then they have maximum mutual information
- mutual information does not have a maximum value, but we will define a normalized version of it that ranges between 0 and 1

MUTUAL INFORMATION $I(X,Y)$ $X \in \{x_1,..,x_m\}$ $Y \in \{y_1,..,y_n\}$

$$H(X) = -\sum_{j=1}^{m} P(X = x_j)\, log_2 P(X = x_j)$$

ENTROPY OF X $H(X) \in [0, log_2 m]$

$$H(Y) = -\sum_{k=1}^{n} P(Y = y_k)\, log_2 P(Y = y_k)$$

ENTROPY OF Y $H(Y) \in [0, log_2 n]$

$$H(X,Y) = -\sum_{j=1}^{m}\sum_{k=1}^{n} P(X = x_j, Y = y_k)\, log_2 P(X = x_j, Y = y_k)$$

JOINT ENTROPY OF X and Y $H(X,Y) \in [0, log_2 mn]$

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

MUTUAL INFORMATION

$$\max I(X,Y) = log_2(min(m,n))$$

$$H(X,Y) = H(Y,X) \Rightarrow I(X,Y) = I(Y,X)$$

IS SYMMETRIC

(*) $0\ log_2 0 = 0$, by convention.

Student Status	Count	p	$-p\log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

X

Grade	Count	p	$-p\log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Y

Student Status	Grade	Count	p	$-p\log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

$$H(X) = -0.45\ log_2\ 0.45 - 0.55\ log_2\ 0.55 = 0.9928$$

$$H(Y) = -0.35\ log_2\ 0.35 - 0.50\ log_2\ 0.50 - 0.15\ log_2\ 0.15 = 1.4406$$

$$H(X,Y) = -0.05\ log_2\ 0.05 - 0.3\ log_2\ 0.3 - 0.1\ log_2\ 0.1 - 0.3\ log_2\ 0.3 - 0.2\ log_2\ 0.2 - 0.05\ log_2\ 0.05 = 2.2710$$

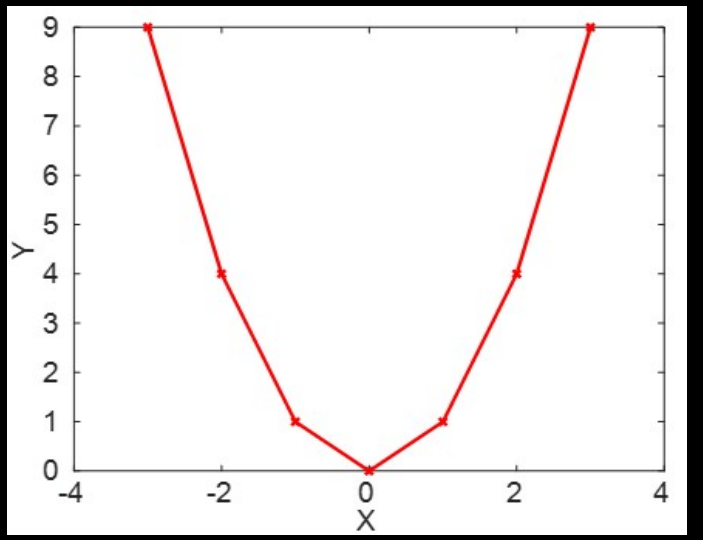
$$I(X,Y) = 0.9928 + 1.4406 - 2{,}2710 = 0.1624 \qquad \max I(X,Y) = log_2\big(min(2,3)\big) = 1$$

$$normalized\ I(X,Y) = \frac{0.1624}{1} = 0.1624$$

(*) in this case there is no change

Mutual Information for nonlinear relationship

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



$$H(X) = -7\left(\frac{1}{7}\,log_2\frac{1}{7}\right) = 2.8071$$

$$H(Y) = -3\left(\frac{2}{7}\,log_2\frac{2}{7}\right) - \left(\frac{1}{7}\,log_2\frac{1}{7}\right) = 1.9502$$

$$H(X,Y) = -7\left(\frac{1}{7}\,log_2\frac{1}{7}\right) = 2.8071$$

$$I(X,Y) = 2.8071 + 1.9502 - 2.8071 = 1.9502$$

$$\max I(X,Y) = log_2\big(min(7,4)\big) = 2$$

$$normalized\ I(X,Y) = \frac{1.9502}{2} = 0.9751$$

(*) X *and* Y are strongly related

$$Corr(X,Y) = 0$$

(**) X *and* Y are unrelated

Sometimes attributes are of many different types, but an overall similarity is needed.

1: For the k^{th} attribute, compute a similarity, $s_k(\mathbf{x}, \mathbf{y})$, in the range [0,1]

2: Define an indicator variable, δ_k, for the k^{th} attribute as follows:

δ_k = 0 if the k^{th} attribute is an asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the k^{th} attribute

δ_k = 1 otherwise

3. Compute

$$similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

May not want to treat all attributes the same.

- Use non-negative weights w_k to compute

$$similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

Can also define a weighted form of distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

RECAP

- SIMILARITY – DISSIMILARITY
- EUCLIDEAN – MINKOWSKI – CITY BLOCK – MAHALANOBIS
- SIMPLE MATCHING COEFFICIENT – JACCARD – COSINE SIMILARITY
- CORRELATION
- ENTROPY – MUTUAL INFORMATION