Data Mining

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Data Mining

Similarity and Dissimilarity Measures

- Outline
  - Similarity and Dissimilarity between Simple Attributes
  - Dissimilarities between Data Objects
  - Similarities between Data Objects
  - Examples of Proximity
  - Mutual Information
  - Issues in Proximity
  - Selecting the Right Proximity Measure

Similarity and Dissimilarity Measures

- Similarity and dissimilarity are important because they are used by a number of data mining techniques, such as clustering, nearest neighbor classification, and anomaly detection.
- In many cases, the initial data set is not needed once these similarities or dissimilarities have been computed.
- Such approaches can be viewed as transforming the data to a similarity (dissimilarity) space and then performing the analysis.

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
  - The term distance is used as a synonym for dissimilarity
- Proximity refers to a similarity or dissimilarity

Transformations

- often applied to convert a similarity to a dissimilarity, or vice versa, or to transform a proximity measure to fall within a particular range, such as [0,1].
  - For instance, we may have similarities that range from 1 to 10, but the particular algorithm or software package that we want to use may be designed to work only with dissimilarities, or it may work only with similarities in the interval [0,1]
- Frequently, proximity measures, especially similarities, are defined or transformed to have values in the interval [0,1].

Transformations

- often applied to convert a similarity to a dissimilarity, or vice versa, or to transform a proximity measure to fall within a particular range, such as [0,1].
  - For instance, we may have similarities that range from 1 to 10, but the particular algorithm or software package that we want to use may be designed to work only with dissimilarities, or it may work only with similarities in the interval [0,1]
- Frequently, proximity measures, especially similarities, are defined or transformed to have values in the interval [0,1].

Transformations

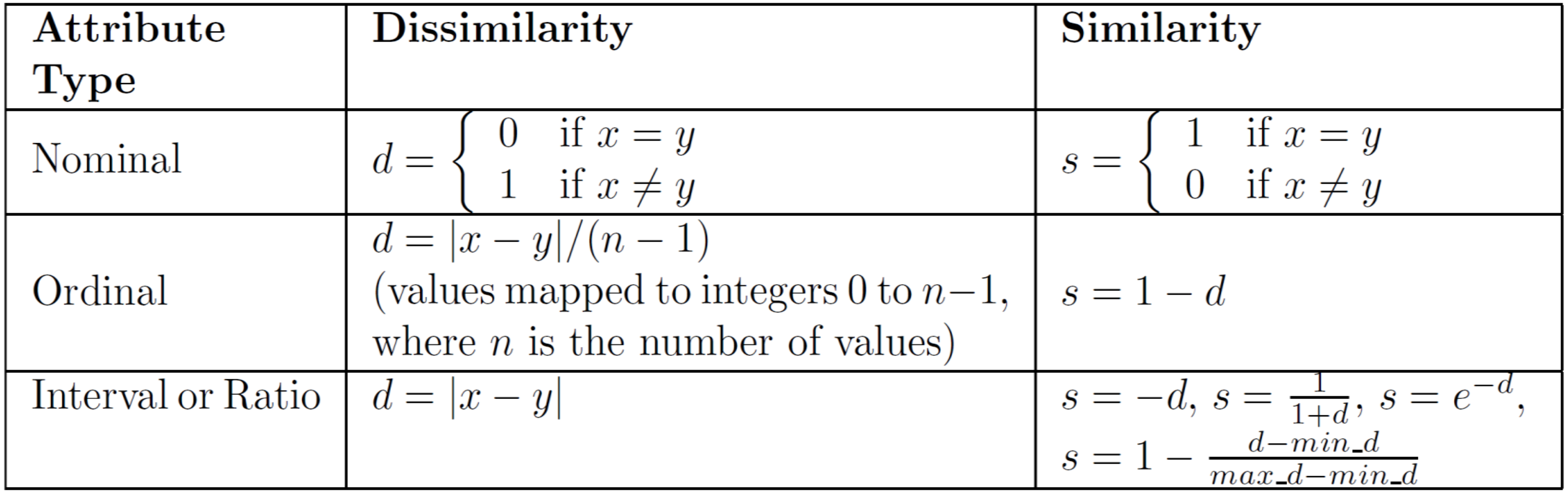
- Example:
  - If the similarities between objects range from 1 (not at all similar) to 10 (completely similar), we can make them fall within the range [0, 1] by using the transformation $s'=(s-1)/9$,where s and s' are the original and new similarity values, respectively.
- The transformation of similarities and dissimilarities to the interval [0, 1]
  - $s'=(s-s_{\text{min}})/(s_{\text{max}}- s_{\text{min}})$,where s_{max} and s_{min} are the maximum and minimum similarity values.
  - $d'=(d-d_{\text{min}})/(d_{\text{max}}- d_{\text{min}})$,where d_{max} and d_{min} are the maximum and minimum dissimilarity values.

Transformations

- However, there can be complications in mapping proximity measures to the interval [0, 1] using a linear transformation.
  - If, for example, the proximity measure originally takes values in the interval $[0,\infty]$, then $d_{\max}$ is not defined and a nonlinear transformation is needed.
  - Values will not have the same relationship to one another on the new scale.
- Consider the transformation $d=d/(1+d)$ for a dissimilarity measure that ranges from 0 to ∞.
  - Given dissimilarities 0, 0.5, 2, 10, 100, 1000
  - Transformed dissimilarities 0, 0.33, 0.67, 0.90, 0.99, 0.999.
- Larger values on the original dissimilarity scale are compressed into the range of values near 1, but whether this is desirable depends on the application.

Similarity/Dissimilarity for Simple Attributes

- 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	$d = \lvert x - y \rvert$	$s = -d,\ s = \frac{1}{1+d},\ s = e^{-d},$ $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

- Next, we consider more complicated measures of proximity between objects that involve multiple attributes:
  - dissimilarities between data objects
  - similarities between data objects.

Distances - Euclidean Distance

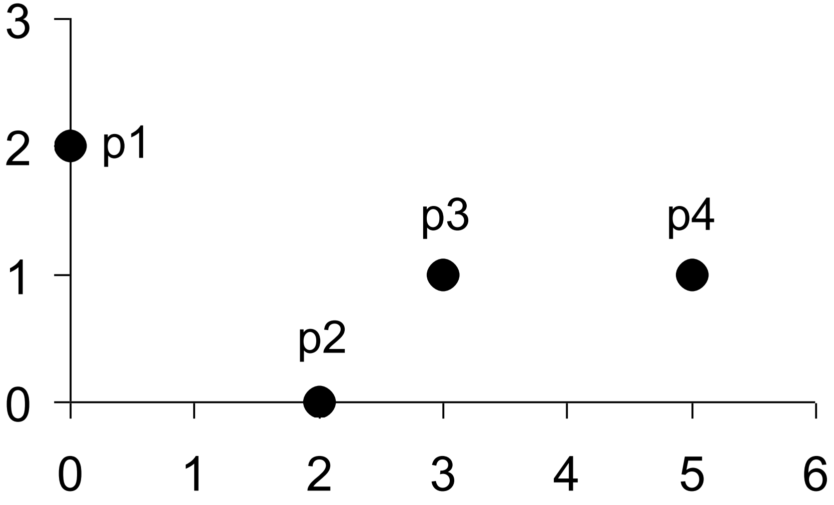
- The Euclidean distance, *d* , between two points, *x* and *y* , in one-, two-, three-, or higher-dimensional space, is given by

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

  – 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.

Distances - Euclidean Distance



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

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

	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

Distances - Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance, and is given by

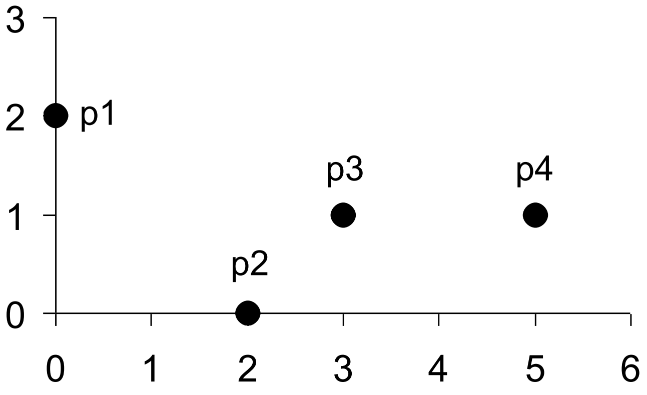
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

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

Distances - Minkowski Distance

- The following are the three most common examples of Minkowski distances.
  - $r = 1$, City block (Manhattan, taxicab, L_1 norm) distance.
    - 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
  - $r = 2$, Euclidean distance (L_2 norm)
  - $r = \infty$, 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.

Distances - Minkowski Distance



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

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

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

Distances - Mahalanobis Distance

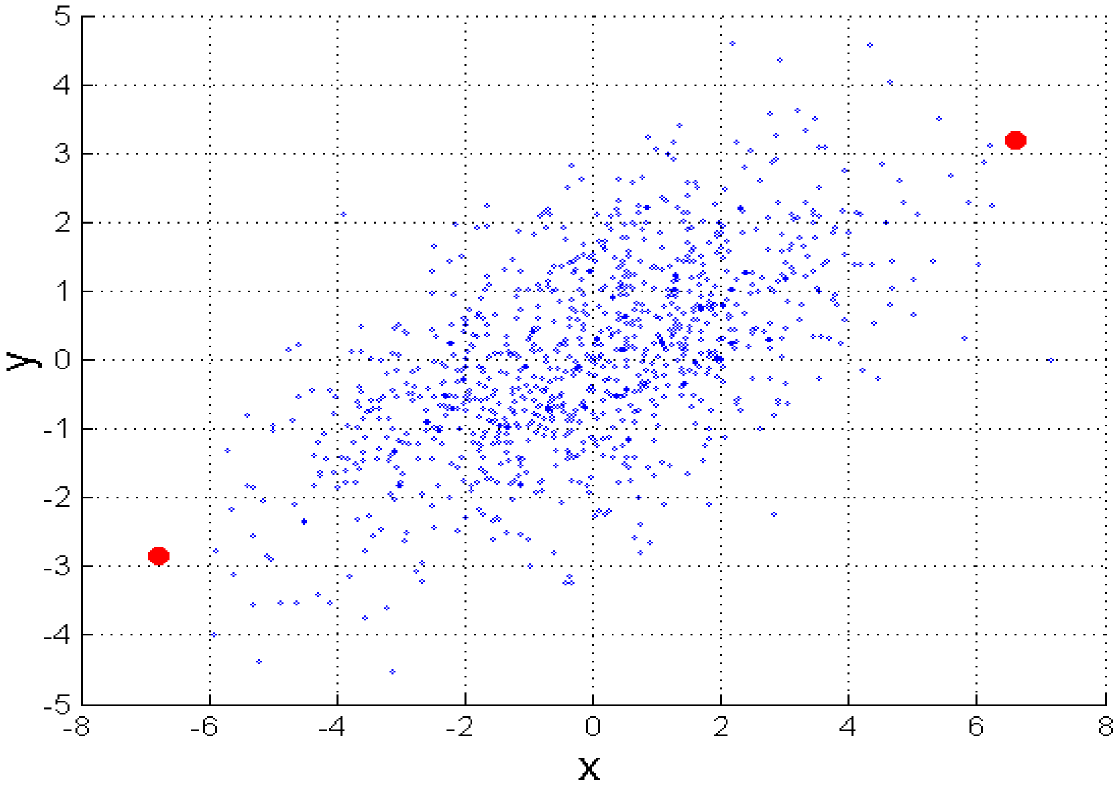
- Mahalonobis distance is the distance between a point and a distribution (not between two distinct points).
  - – It is effectively a multivariate equivalent of the Euclidean distance.
    - It transforms the columns into uncorrelated variables
    - Scale the columns to make their variance equal to 1
    - Finally, it calculates the Euclidean distance.
- It is defined as

$$\mathrm{Mahalanobis}(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\Sigma^{-1}(\mathbf{x}-\mathbf{y})}$$

  - – where Σ^{-1} is the inverse of the covariance matrix of the data.

Distances - Mahalanobis Distance

- In the Figure, there are 1000 points, whose *x* and *y* attributes have a correlation of 0.6.



  - The Euclidean distance between the two large points at the opposite ends of the long axis of the ellipse is 14.7, but Mahalanobis distance is only 6.
    - This is because the Mahalanobis distance gives less emphasis to the direction of largest variance.

Distances - Mahalanobis Distance

- Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

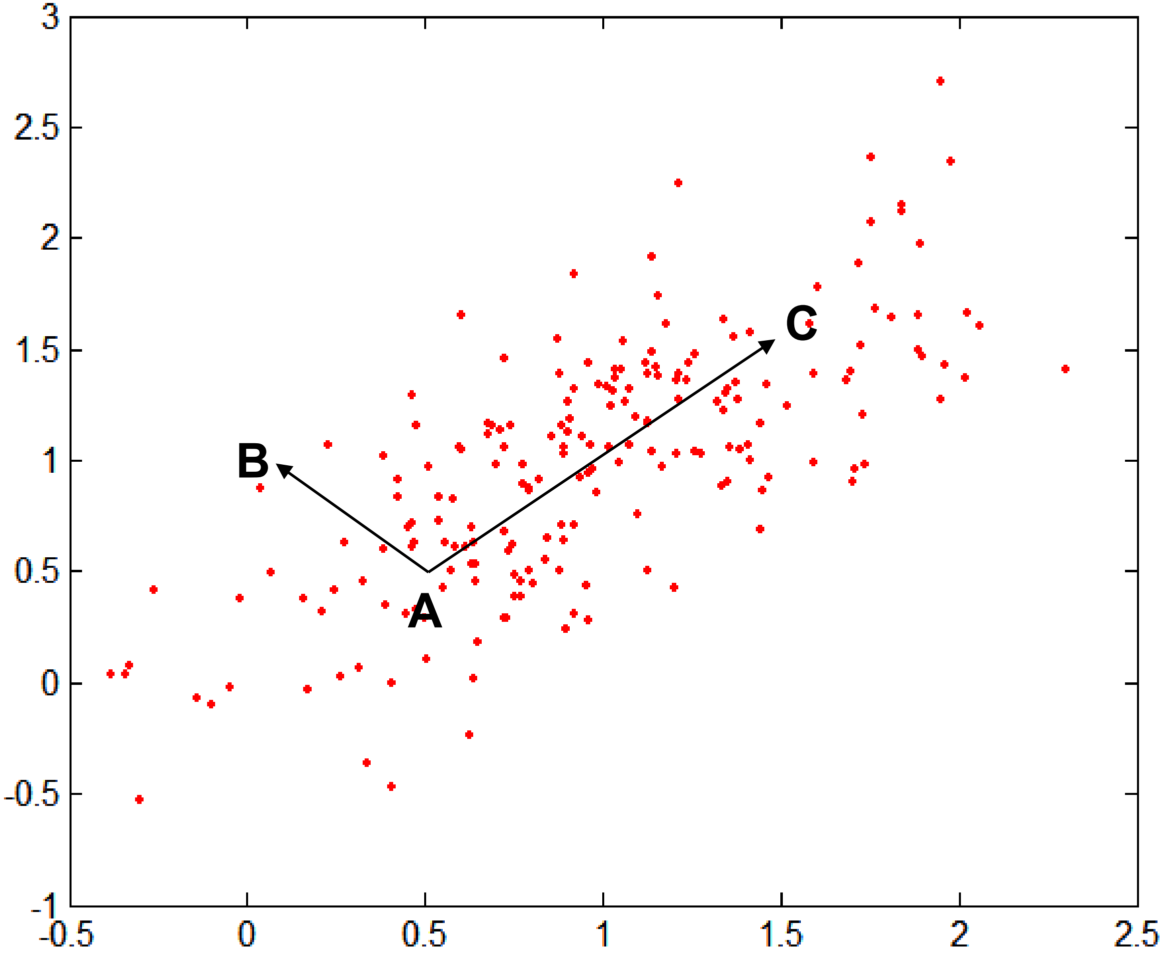
A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4



Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well-known properties.
- If $d(x, y)$ is the distance between two points, x and y, then the following properties hold.
  - Positivity
    - $d(x, y) \geq 0$ for all x and y
    - $d(x, y) = 0$ only if $x = y$
  - Symmetry
    - $d(x, y) = d(y, x)$ for all x and y
  - Triangle Inequality
    - $d(x, z) \leq d(x, y) + d(y, z)$ for all points x, y, and z
- Measures that satisfy all three properties are known as metrics

Common Properties of a Similarity

- If $s(x, y)$ is the similarity between points x and y, then the typical properties of similarities are the following:
  - Positivity
    - $s(x, y) = 1$ only if $x = y$. $(0 \leq s \leq 1)$
  - Symmetry
    - $s(x, y) = s(y, x)$ for all x and y
- For similarities, the triangle inequality typically does not hold
  - However, a similarity measure can be converted to a metric distance

A Non-symmetric Similarity Measure Example

- Consider an experiment in which people are asked to classify a small set of characters as they flash on a screen.
  - The confusion matrix for this experiment records how often each character is classified as itself, and how often each is classified as another character.
  - Using the confusion matrix, we can define a similarity measure between a character x and a character y as the number of times that x is misclassified as y,
    - but note that this measure is not symmetric.

A Non-symmetric Similarity Measure Example

- For example, suppose that “0” appeared 200 times and was classified as a “0” 160 times, but as an “o” 40 times.
- Likewise, suppose that “o” appeared 200 times and was classified as an “o” 170 times, but as “0” only 30 times.
  - Then, s(0,o) = 40, but s(o, 0) = 30.
- In such situations, the similarity measure can be made symmetric by setting
  - $s'(x, y) = s'(y, x) = (s(x, y)+s(y, x))/2$,
    - where *s* indicates the new similarity measure.

Similarity Measures for Binary Data

- Similarity measures between objects that contain only binary attributes are called similarity coefficients, and typically have values between 0 and 1.
- Let x and y be two objects that consist of *n* binary attributes.
  - – The comparison of two binary vectors, leads to the following quantities (frequencies):
    - f_{00} = the number of attributes where x is 0 and y is 0
    - f_{01} = the number of attributes where x is 0 and y is 1
    - f_{10} = the number of attributes where x is 1 and y is 0
    - f_{11} = the number of attributes where x is 1 and y is 1

Similarity Measures for Binary Data

- Simple Matching Coefficient (SMC)
  - One commonly used similarity coefficient

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}} = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

  - This measure counts both presences and absences equally.
    - Consequently, the SMC could be used to find students who had answered questions similarly on a test that consisted only of true/false questions.

Similarity Measures for Binary Data

- Jaccard Similarity Coefficient
  - frequently used to handle objects consisting of asymmetric binary attributes

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

  - This measure counts both presences and absences equally.
    - Consequently, the SMC could be used to find students who had answered questions similarly on a test that consisted only of true/false questions.

SMC versus Jaccard: Example

- Calculate SMC and J for the binary vectors,

  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 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)

-

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

  $= (0 + 7) / (2 + 1 + 0 + 7) \qquad = 0.7$

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

  $= 0 / (2 + 1 + 0) \qquad = 0$

Cosine Similarity

- Cosine Similarity is one of the most common measures of document similarity
- If x and y are two document vectors, then

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{\mathbf{x}'\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|}$$

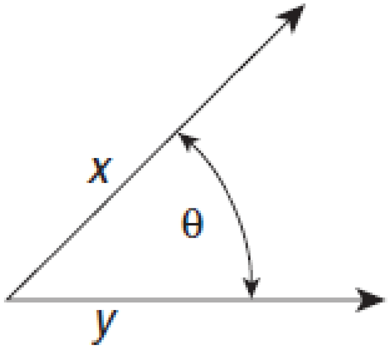
  - where ′ indicates vector or matrix transpose and ⟨x,y⟩ indicates the inner product of the two vectors,

$\langle \mathbf{x},\mathbf{y}\rangle = \sum_{k=1}^{n} x_k y_k = \mathbf{x}'\mathbf{y}$ and $\|x\|$ is the length of vector x,

$$\|\mathbf{x}\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\langle \mathbf{x},\mathbf{x}\rangle} = \sqrt{\mathbf{x}'\mathbf{x}}$$

Cosine Similarity

- Cosine similarity really is a measure of the (cosine of the) angle between x and y.



  - Thus, if the cosine similarity is 1, the angle between x and y is 0°, and x and y are the same except for length.
  -
  - If the cosine similarity is 0, then the angle between x and y is 90°, and they do not share any terms (words).

- It can also be written as

$$\cos(\mathbf{x},\mathbf{y}) = \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle = \langle \mathbf{x}', \mathbf{y}' \rangle$$

Cosine Similarity - Example

- Cosine Similarity between two document vectors
- This example calculates the cosine similarity for the following two data objects, which might represent document vectors:

$$x = (3, 2, 0, 5, 0, 0, 0, 2, 0, 0)$$

$$y = (1, 0, 0, 0, 0, 0, 0, 1, 0, 2)$$

$$\langle x,y \rangle = 3 \times 1 + 2 \times 0 + 0 \times 0 + 5 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 2 = 5$$

$$\|x\| = \sqrt{3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2} = 6.48$$

$$\|y\| = \sqrt{1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2} = 2.45$$

$$\cos(x, y) = \frac{\langle x,y \rangle}{\|x\| \times \|y\|} = \frac{5}{6.48 \times 2.45} = 0.31$$

Extended Jaccard Coefficient

- Also known as Tanimoto Coefficient
- The extended Jaccard coefficient can be used for document data and that reduces to the Jaccard coefficient in the case of binary attributes.
- This coefficient, which we shall represent as EJ, is defined by the following equation:

$$EJ(\mathrm{x},\mathrm{y}) = \frac{\langle \mathrm{x},\mathrm{y} \rangle}{\|\mathrm{x}\|^2 + \|\mathrm{y}\|^2 - \langle \mathrm{x},\mathrm{y} \rangle} = \frac{\mathrm{x}'\mathrm{y}}{\|\mathrm{x}\|^2 + \|\mathrm{y}\|^2 - \mathrm{x}'\mathrm{y}}$$

Correlation

- used to measure the linear relationship between two sets of values that are observed together.
  - Thus, correlation can measure the relationship between two variables (height and weight) or between two objects (a pair of temperature time series).
- Correlation is used much more frequently to measure the similarity between attributes
  - since the values in two data objects come from different attributes, which can have very different attribute types and scales.
- There are many types of correlation

Correlation - Pearson’s correlation

- between two sets of numerical values, i.e., two vectors, x and y, is defined by:

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

  – where the following standard statistical notation and definitions are used:

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

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

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

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

Correlation – Example (Perfect Correlation)

- Correlation is always in the range −1 to 1.
  - A correlation of 1 (−1) means that x and y have a perfect positive (negative) linear relationship;
    - that is, $x_k = a y_k + b$, where a and b are constants.
- The following two vectors x and y illustrate cases where the correlation is −1 and +1, respectively.

x = (−3, 6, 0, 3,−6) x = (3, 6, 0, 3, 6)

y = (1,−2, 0,−1, 2) y = (1, 2, 0, 1, 2)

corr(x, y) = −1 $x_k = -3y_k$ corr(x, y) = 1 $x_k = 3y_k$

Correlation – Example (Nonlinear Relationships)

- If the correlation is 0, then there is no linear relationship between the two sets of values.
  - – However, nonlinear relationships can still exist.
    - In the following example, $y_k = x_k^2$, but their correlation is 0.

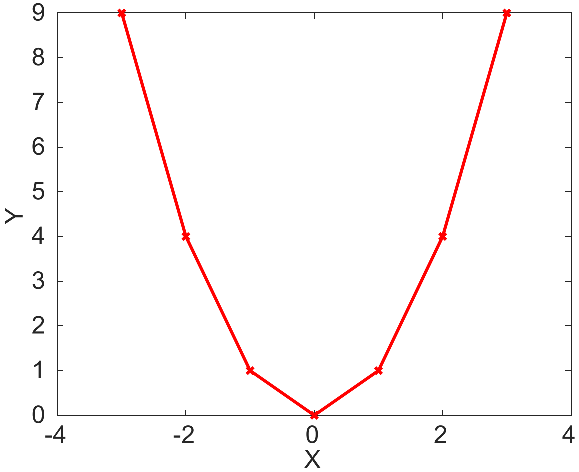
x = (-3, -2, -1, 0, 1, 2, 3)

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

$y_k = x_k^2$

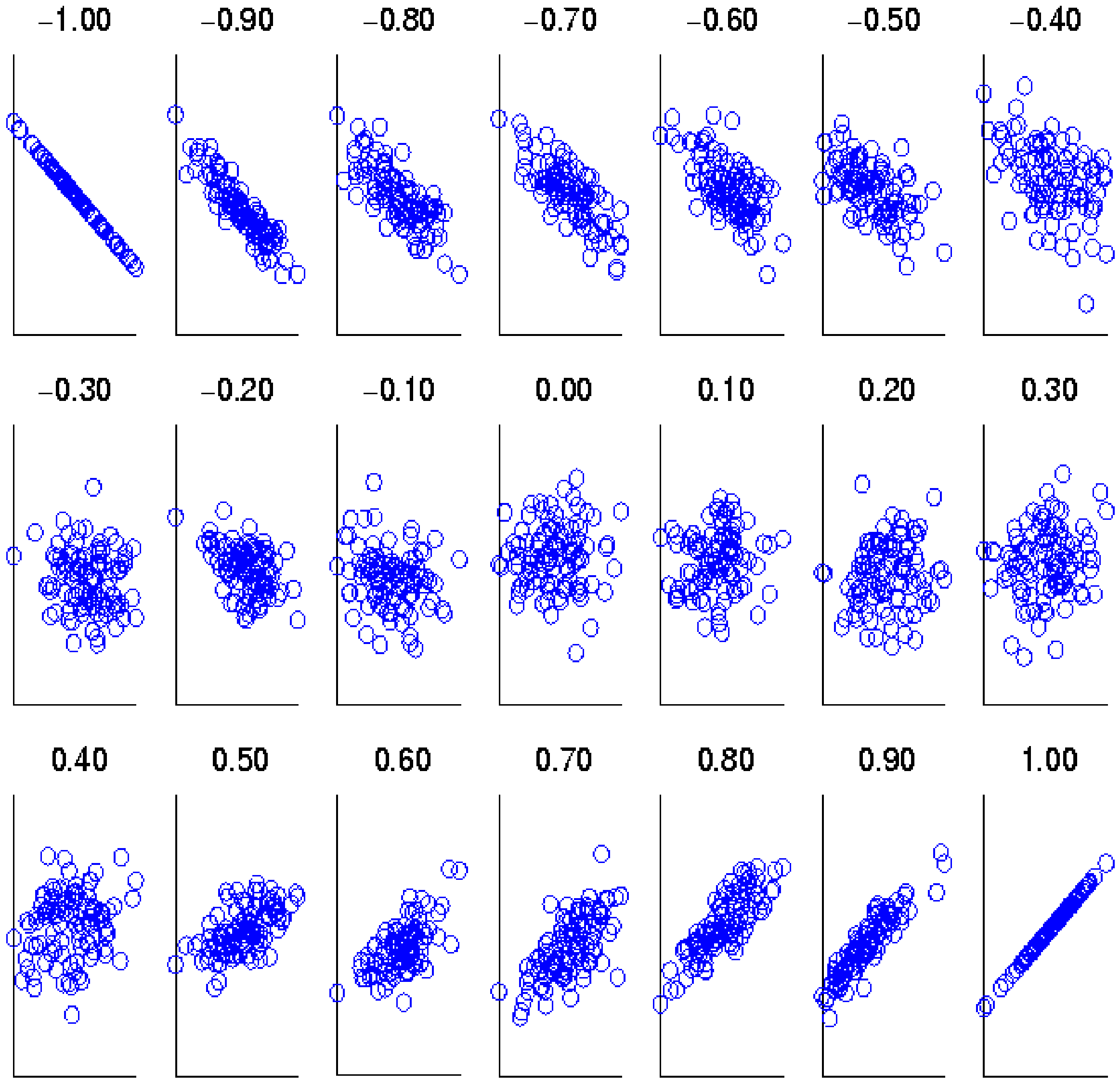
mean(x) = 0, mean(y) = 4

std(x) = 2.16, std(y) = 3.74



$$corr = \frac{(-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+(3)(5)}{6 \times 2.16 \times 3.74} = 0$$

Visually Evaluating Correlation



- Scatter plots showing the similarity from –1 to 1.

Correlation vs Cosine vs Euclidean Distance

- 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

- Consider the example
  - $x = (1, 2, 4, 3, 0, 0, 0)$, $y = (1, 2, 3, 4, 0, 0, 0)$
  - $y_s = y \times 2 = (2, 4, 6, 8, 0, 0, 0)$ $y_t = y + 5 = (6, 7, 8, 9, 5, 5, 5)$

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

Correlation vs cosine vs Euclidean distance

- 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.

Comparison of Proximity Measures

- 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 Based Measures

- Information theory is a well-developed and fundamental disciple 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), $x_1, x_2 \ldots, x_n$
  - 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
  - Thus, entropy is a measure of how many bits it takes to represent an observation of *X* on average

Entropy Examples

- 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

- What is the entropy of a fair four-sided die?

Entropy for Sample Data: Example

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

- Maximum entropy is $\log_2 5 = 2.3219$

Entropy for Sample Data

- 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

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

- For continuous data, the calculation is harder

Mutual Information

- used as a measure of similarity between two sets of paired values that is sometimes used as an alternative to correlation, particularly when a nonlinear relationship is suspected between the pairs of values.
  - This measure comes from information theory, which is the study of how to formally define and quantify information.
  - It is 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 0.

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
- For discrete variables, this is easy to compute
- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

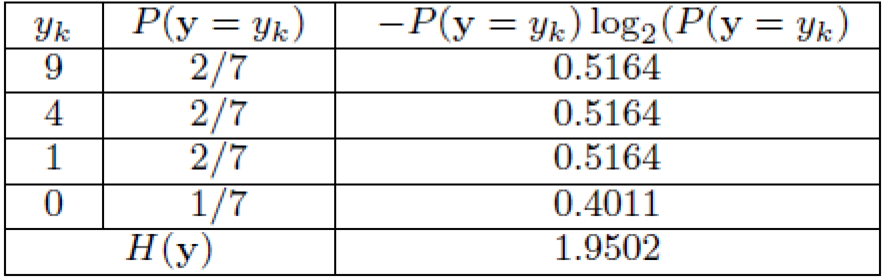
Mutual Information Example

- Evaluating Nonlinear Relationships with Mutual Information
  - Recall Example where $y_k = x_k^2$, but their correlation was 0.

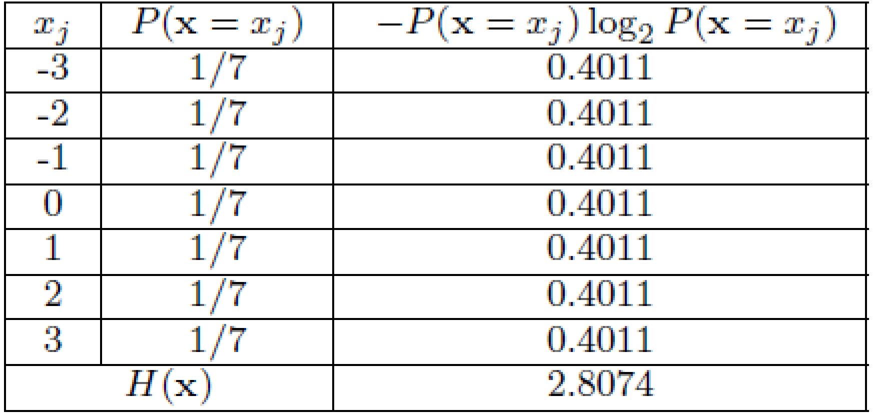
x = (−3,−2,−1, 0, 1, 2, 3) y = (9, 4, 1, 0, 1, 4, 9)

$I(x, y) = H(x) + H(y) - H(x, y) = 1.9502$

Entropy for y



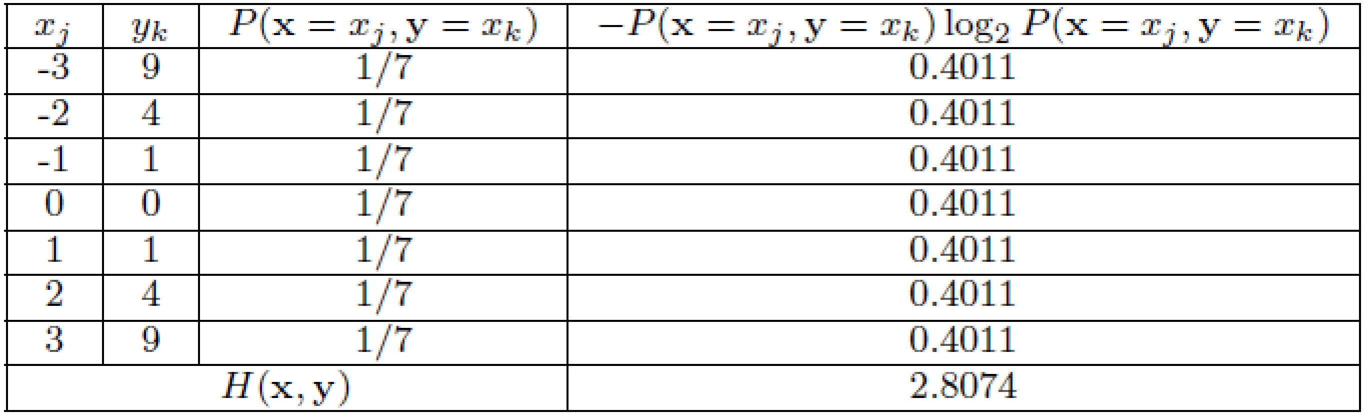
y_k	$P(\mathrm{y} = y_k)$	$-P(\mathrm{y} = y_k)\log_2(P(\mathrm{y} = y_k))$
9	2/7	0.5164
4	2/7	0.5164
1	2/7	0.5164
0	1/7	0.4011
$H(\mathrm{y})$		1.9502



x_j	$P(\mathrm{x} = x_j)$	$-P(\mathrm{x} = x_j)\log_2 P(\mathrm{x} = x_j)$
-3	1/7	0.4011
-2	1/7	0.4011
-1	1/7	0.4011
0	1/7	0.4011
1	1/7	0.4011
2	1/7	0.4011
3	1/7	0.4011
$H(\mathrm{x})$		2.8074

Entropy for x

Joint entropy for **x** and **y**



x_j	y_k	$P(\mathrm{x} = x_j, \mathrm{y} = x_k)$	$-P(\mathrm{x} = x_j, \mathrm{y} = x_k)\log_2 P(\mathrm{x} = x_j, \mathrm{y} = x_k)$
-3	9	1/7	0.4011
-2	4	1/7	0.4011
-1	1	1/7	0.4011
0	0	1/7	0.4011
1	1	1/7	0.4011
2	4	1/7	0.4011
3	9	1/7	0.4011
$H(\mathrm{x}, \mathrm{y})$			2.8074

Mutual Information Example

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

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

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

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

Maximal Information Coefficient

- Applies mutual information to two continuous variables
- Consider the possible binnings of the variables into discrete categories
  - $n_X \times n_Y \leq N^{0.6}$ where
    - n_X is the number of values of X
    - n_Y is the number of values of Y
    - N is the number of samples (observations, data objects)
- Compute the mutual information
  - Normalized by $\log_2(\min(n_X, n_Y)$
- Take the highest value

- Reshef, David N., Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. "Detecting novel associations in large data sets." *science* 334, no. 6062 (2011): 1518-1524.

General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.
  - For the k^{th} attribute, compute a similarity, $s_k(x, y)$, in the range [0, 1].
  - Define an indicator variable, δ_k, for the k^{th} attribute as follows:
    - $\delta_k = 0$ 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 kth attribute
    - $\delta_k = 1$ otherwise
  - Compute

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

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use non-negative weights ω_k
  - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$
- Can also define a weighted form of distance

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