School of Computing
UNIVERSITY OF GEORGIA

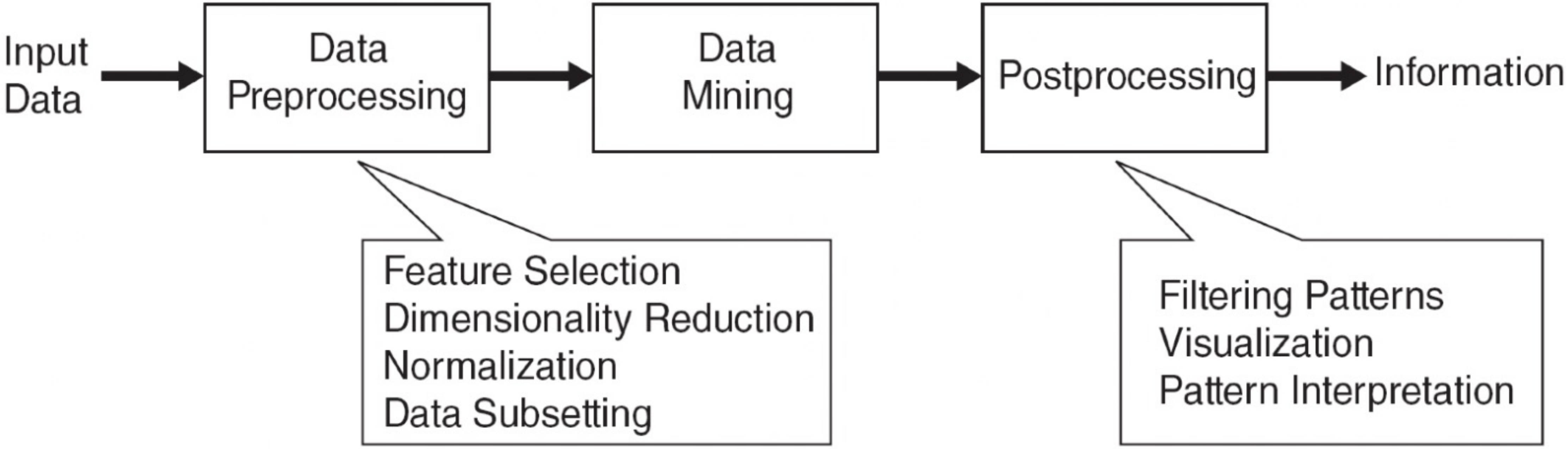
CSCI 4380/6380 Data Mining

Fei Dou

Assistant Professor
School of Computing
University of Georgia

September 14,19, 2023

Recap: Data Mining Process



Recap: Data Preprocessing

- **Data cleaning**:
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Binning, Concept hierarchy generation

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

Data Matrix and Dissimilarity Matrix

- **Data Matrix**: n data points with p dimensions.
- **Dissimilarity matrix**: n data points, but registers only the distance.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \qquad \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

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	$d = \|x - y\|/(n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \|x - y\|$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Proximity Measure for Nominal Attributes

- **Nominal Attributes**: Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)

- **Method 1**: Simple matching
  - m : number of matches, p : total number of variables;
  - $d(i,j) = \frac{p - m}{p}$

- **Method 2**: Use a large number of binary attributes
  - Creating a new binary attribute for each of the M nominal states

Similarity Between Binary Vectors

- 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 and Jaccard Coefficients

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

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

SMC versus Jaccard: Example

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)

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

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

Cosine Similarity

If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = \langle d_1,d_2\rangle / ||d_1||\ ||d_2|| ,$$

where $\langle d_1,d_2\rangle$ indicates inner product or vector dot product of vectors, d_1 and d_2, and $||\ d\ ||$ is the length of vector d.

- Example:

d_1 = 3 2 0 5 0 0 0 2 0 0

d_2 = 1 0 0 0 0 0 0 1 0 2

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

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

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

$\cos(d_1, d_2) = 0.3150$

Euclidean Distance

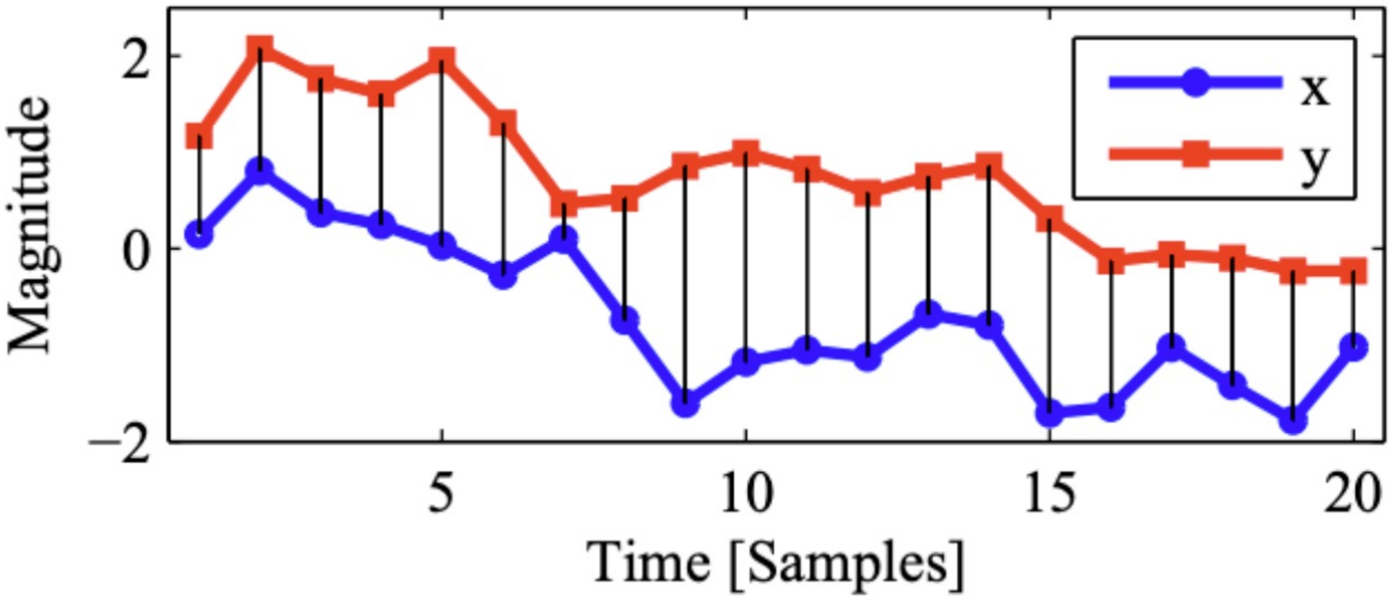
Euclidean Distance

- Two data objects: x = $[x_1, x_2, \cdots, x_n]^\top \in \mathbb{R}^n$ and y = $[y_1, y_2, \cdots, y_n]^\top \in \mathbb{R}^n$
- Z-normalization (standardization) is necessary if scales differ.

$$\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{y}_i)^2}$$

where n is the number of dimensions (attributes) and x_i and y_i are, respectively, the i^{th} attributes (components) or data objects x and y.

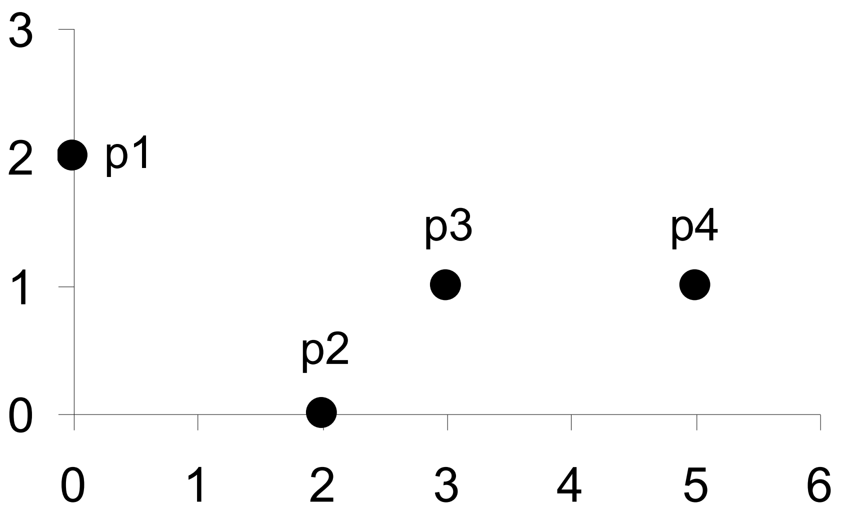
Euclidean Distance



- If $f(\boldsymbol{x}) \in \mathbb{R}^n$ denotes the feature extracted on x and $f_{\boldsymbol{i}}(\boldsymbol{x})$ denotes the i-th dimension.

$$\boldsymbol{d}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{k=1}^{n} (f_i(\boldsymbol{x}) - f_i(\boldsymbol{y}))^2}$$

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

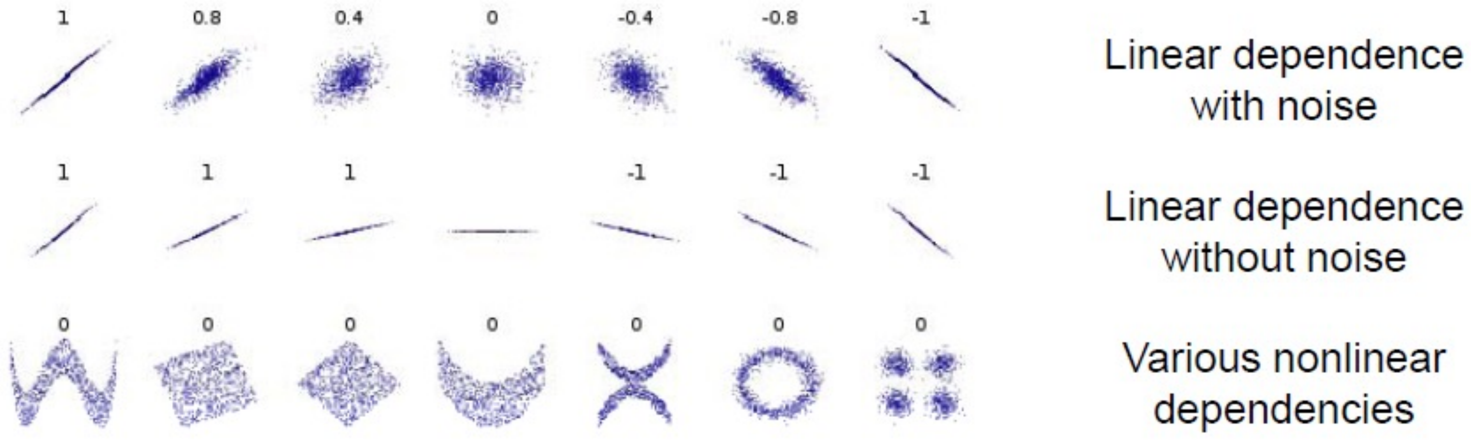
Distance Matrix

Recap: Correlation

- ***Correlation coefficient*** is the covariance normalized by the standard deviations of the two variables

$$\mathrm{corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

  - It is also called Pearson's correlation coefficient and it is denoted $\rho(X, Y)$
  - The values are in the interval $[-1, 1]$
  - It only reflects linear dependence between variables, and it does not measure non-linear dependencies between the variables



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

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = cosine(\boldsymbol{x} - \overline{\boldsymbol{x}}, \boldsymbol{y} - \overline{\boldsymbol{y}},)$$

Correlation vs Cosine vs Euclidean Distance

- Example
  - $x = (1,2,4,3,0,0,0)$, $y = (1,2,3,4,0,0,0)$
  - $y_s = y * 2$ (scaled version of y)
  - $y_t = y + 5$ (translated version)

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?
  - Example 1: Comparing documents using the frequencies of words
  - Example 2: Comparing the temperature in Celsius of two locations
  - Example 3: Comparing two time series of temperature measured in Celsius

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_i and y_i are, respectively, the i^{th} attributes (components) or data objects x and y.

- If r = 1, **City block (Manhattan, $L1$ 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
- If r = 2, Euclidean Distance.
- If $r = \infty$ "supremum" (L_{max} norm, L_∞ norm) distance, i.e., $\max_i |x_i - y_i|$
  - 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 : Examples

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

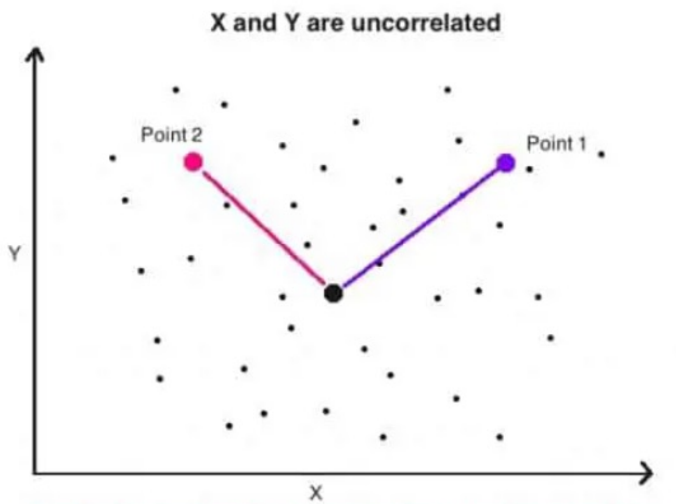
L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

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

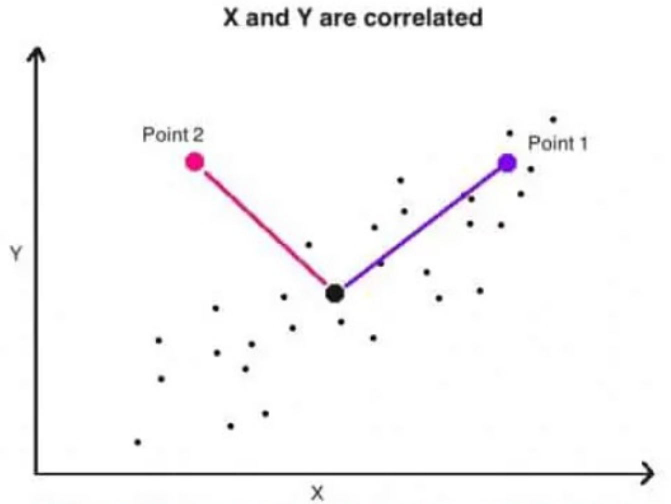
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Mahalanobis Distance

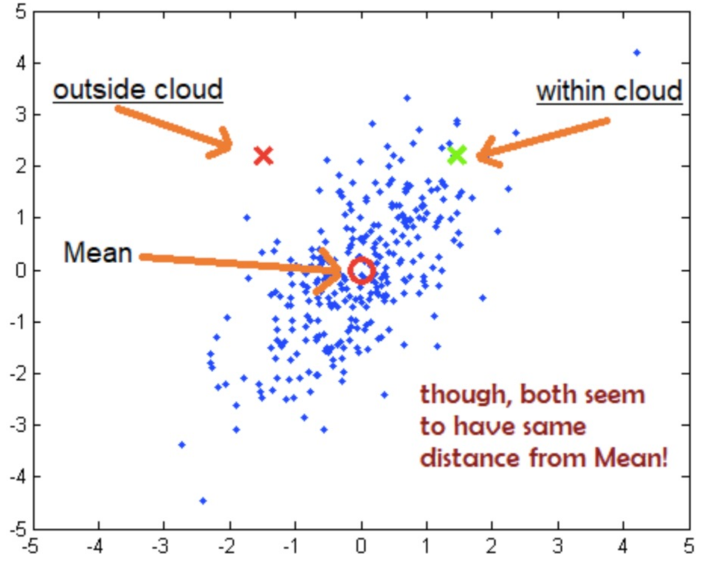


When X and Y are uncorrelated, the Euclidean distance from the Centroid can be useful to infer if a point is member of the distribution. The farther it is, the less likely it is a member.



Both Point 1 and Point 2 have the same Euclidean distance from centroid. But only Point 1 is a member of the distribution. To detect Point 2 as outlier, dist(Point 2, centroid) should be much higher than dist(Point 1, Centroid). Mahalanobis distance can be used here instead.

same Euclidean distance but different Mahalanobis distance



- Mahalanobis Distance is the distance between a point and a distribution, which is defined as:

$$d(\boldsymbol{x}, \overline{\boldsymbol{\mu}}) = \sqrt{(\boldsymbol{x} - \overline{\boldsymbol{\mu}})^T \Sigma^{-1} (\boldsymbol{x} - \overline{\boldsymbol{\mu}})}$$

$\Sigma = \frac{1}{n-1} \sum (\boldsymbol{x} - \overline{\boldsymbol{x}})\,(\boldsymbol{x} - \overline{\boldsymbol{x}})^T$ is the covariance matrix

Common Properties of a Distance

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

- $d(\mathbf{x},\mathbf{y}) \geq 0$ for all x and y, and $d(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all x and y. (Symmetry)
- $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})$ for all **x, y**, and **z** (Triangle Inequality)

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

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

Common Properties of a Similarity

- 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 $\mathbf{y}$.

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 and Probability

- Information relates to possible outcomes of an event, e.g., flip of a coin.
- 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,
    - i.e., 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

$-0.75 \cdot \log_2(0.75) = 0.3113$

$-0.15 \cdot \log2(0.15) = 0.4105$

$-0.05 \cdot \log2(0.05) = 0.2161$

0

$-0.05 \cdot \log2(0.05) = 0.2161$

$H = 0.3113 + 0.4105 + 0.2161$
$+ 0 + 0.2161 = 1.1540$

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

- 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

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