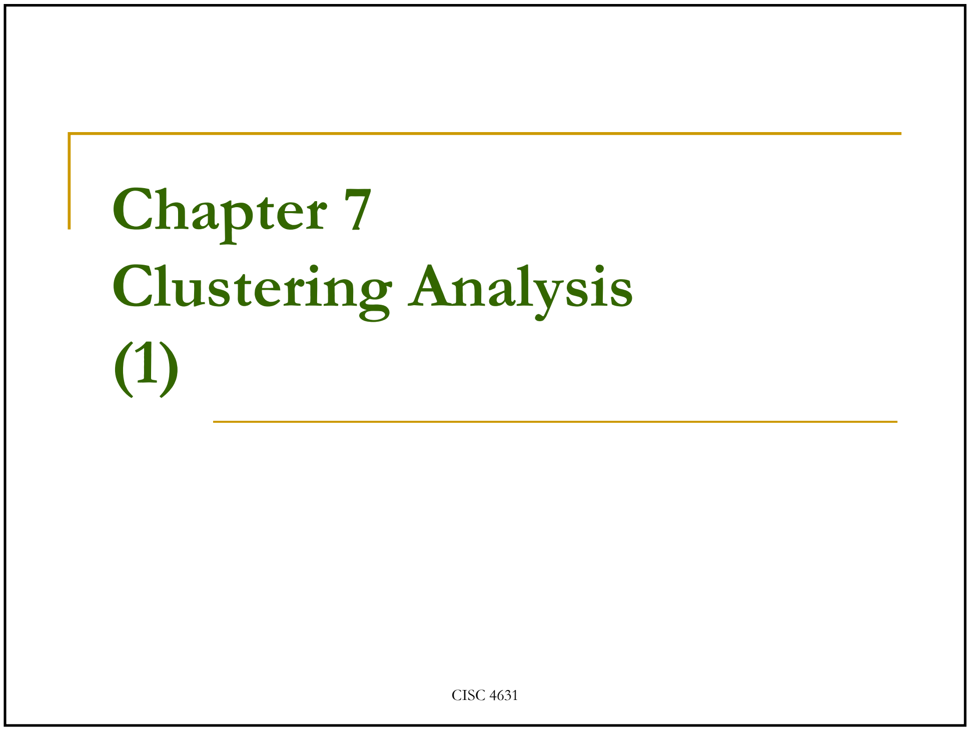
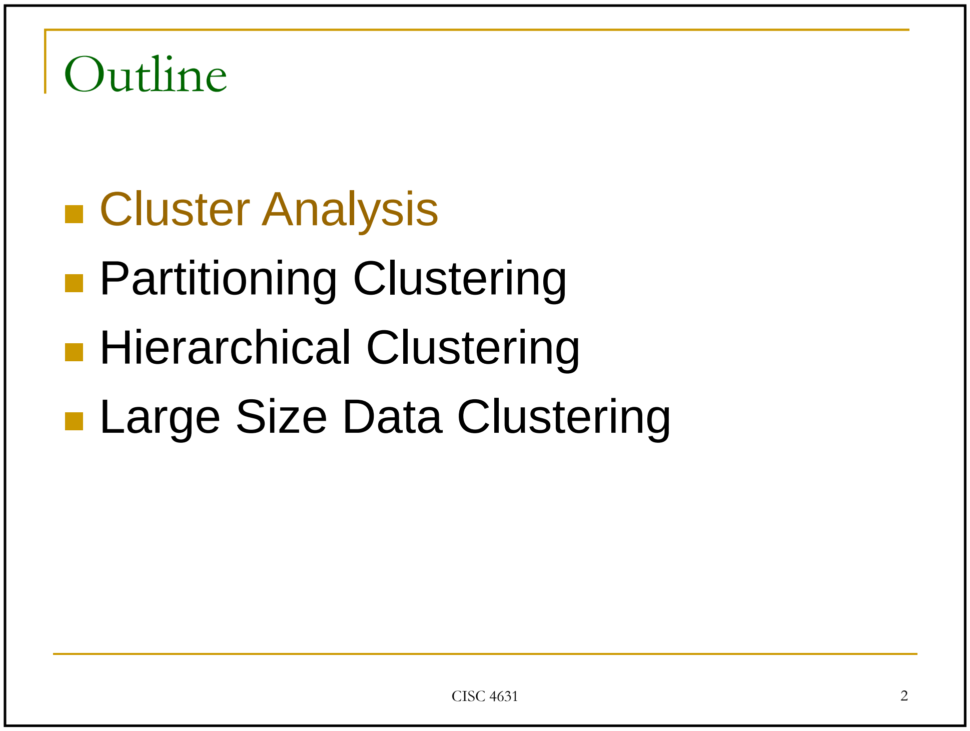
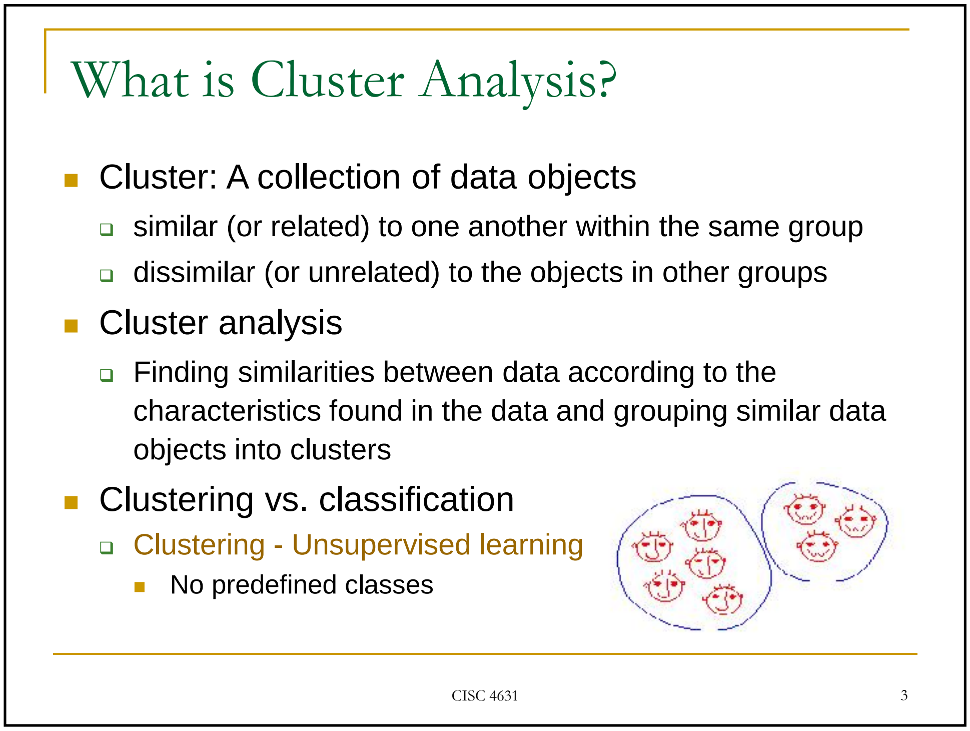


Chapter 7
Clustering Analysis
(1)



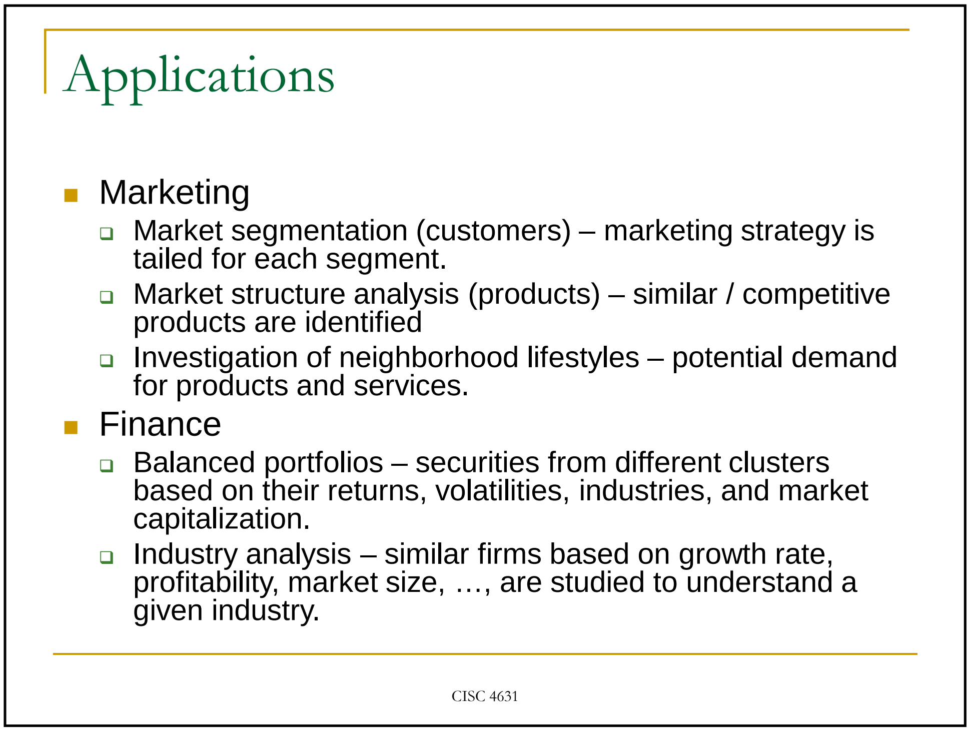
Outline

- Cluster Analysis
- Partitioning Clustering
- Hierarchical Clustering
- Large Size Data Clustering



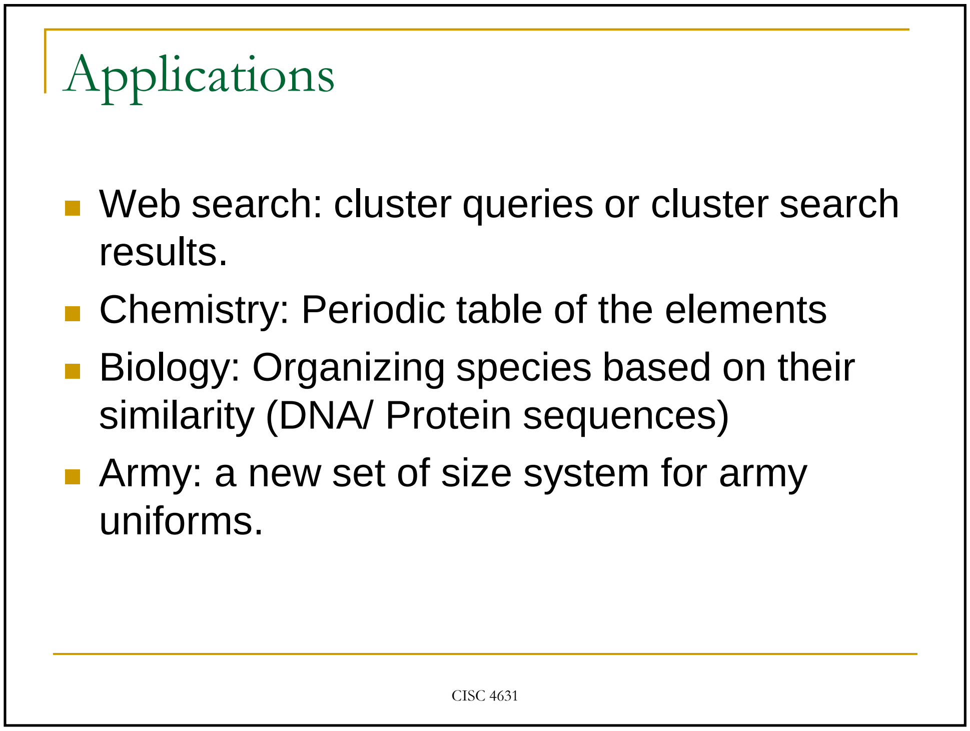
What is Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Clustering vs. classification
  - Clustering - Unsupervised learning
    - No predefined classes



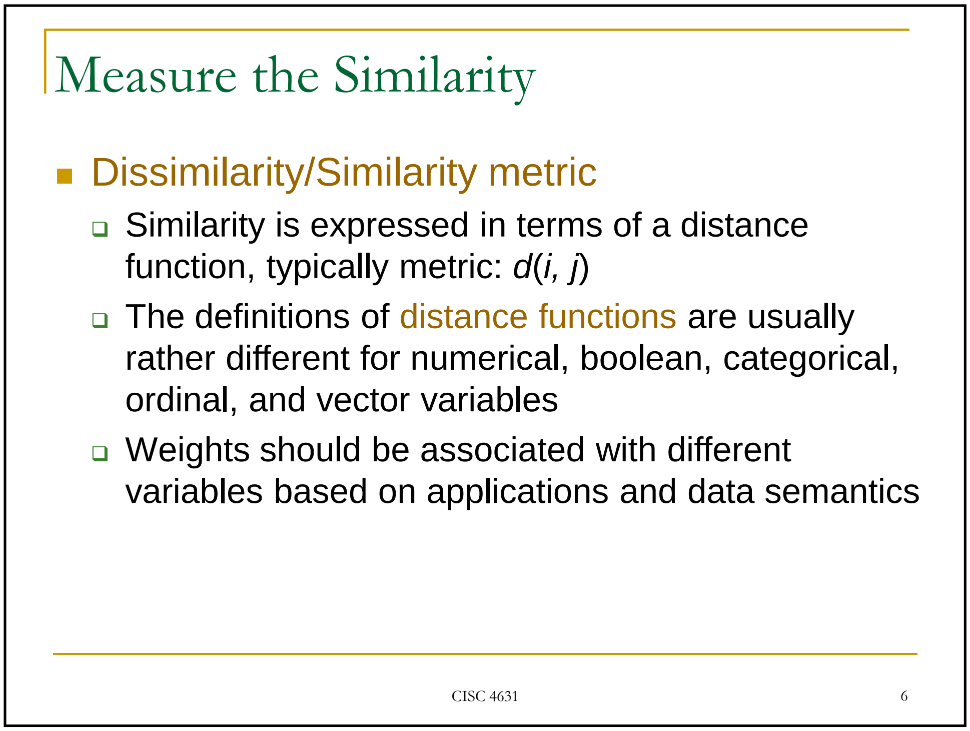
Applications

- Marketing
  - Market segmentation (customers) – marketing strategy is tailed for each segment.
  - Market structure analysis (products) – similar / competitive products are identified
  - Investigation of neighborhood lifestyles – potential demand for products and services.
- Finance
  - Balanced portfolios – securities from different clusters based on their returns, volatilities, industries, and market capitalization.
  - Industry analysis – similar firms based on growth rate, profitability, market size, …, are studied to understand a given industry.



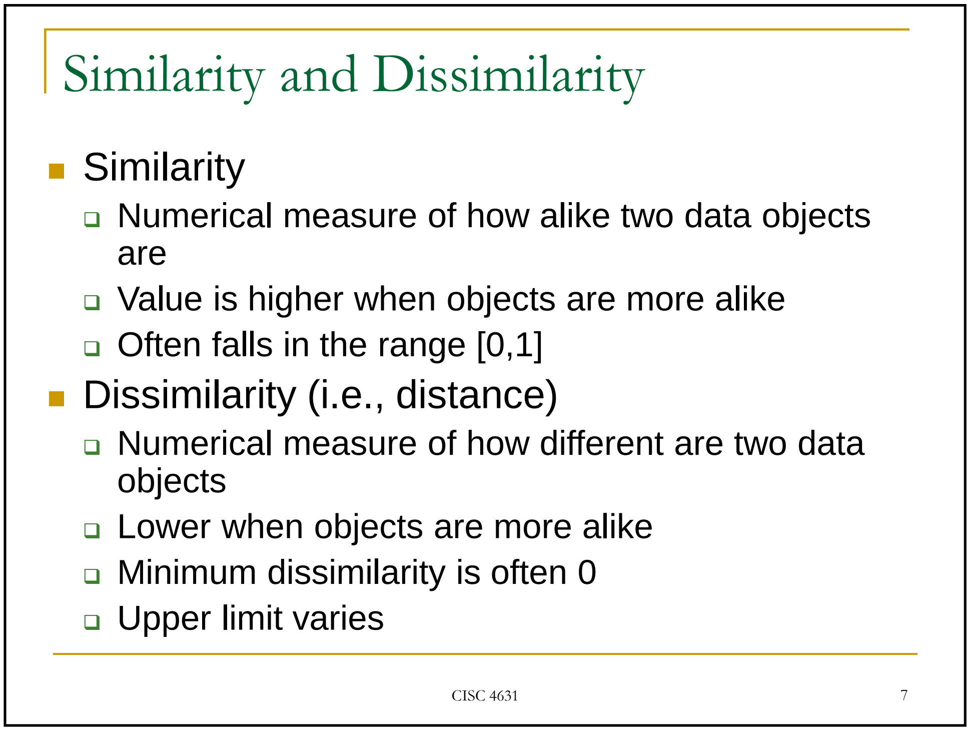
Applications

- Web search: cluster queries or cluster search results.
- Chemistry: Periodic table of the elements
- Biology: Organizing species based on their similarity (DNA/ Protein sequences)
- Army: a new set of size system for army uniforms.



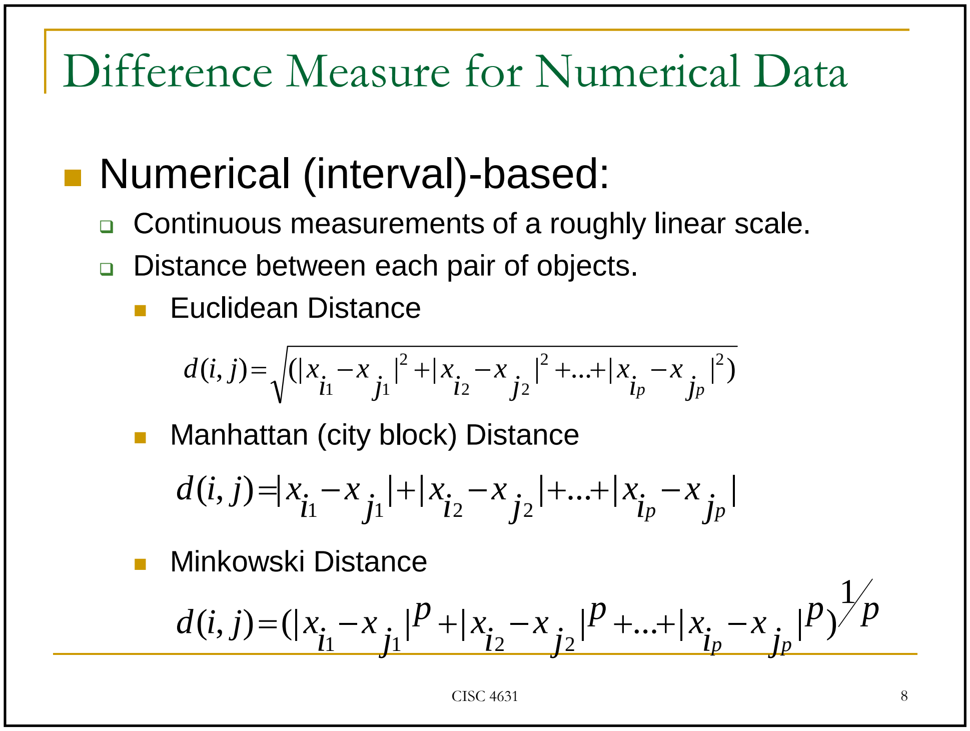
Measure the Similarity

- Dissimilarity/Similarity metric
  - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
  - The definitions of distance functions are usually rather different for numerical, boolean, categorical, ordinal, and vector variables
  - Weights should be associated with different variables based on applications and data semantics



Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- Dissimilarity (i.e., distance)
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies



Difference Measure for Numerical Data

- Numerical (interval)-based:
  - Continuous measurements of a roughly linear scale.
  - Distance between each pair of objects.
    - Euclidean Distance

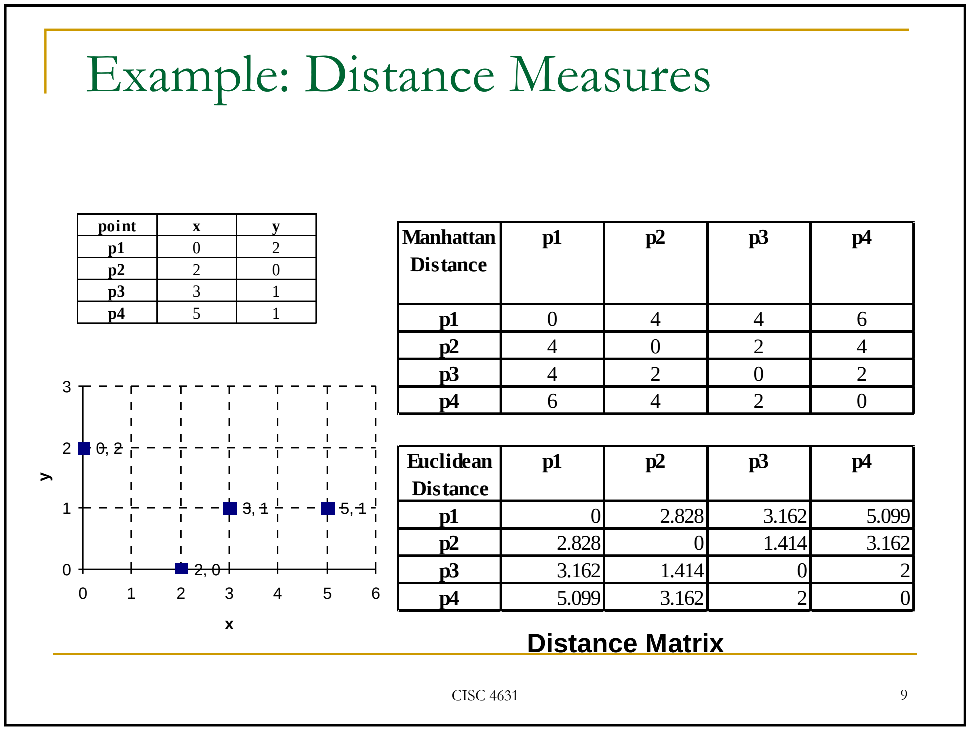
$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$$

    - Manhattan (city block) Distance

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$$

    - Minkowski Distance

$$d(i,j)=(|x_{i1}-x_{j1}|^p+|x_{i2}-x_{j2}|^p+...+|x_{ip}-x_{jp}|^p)^{1/p}$$



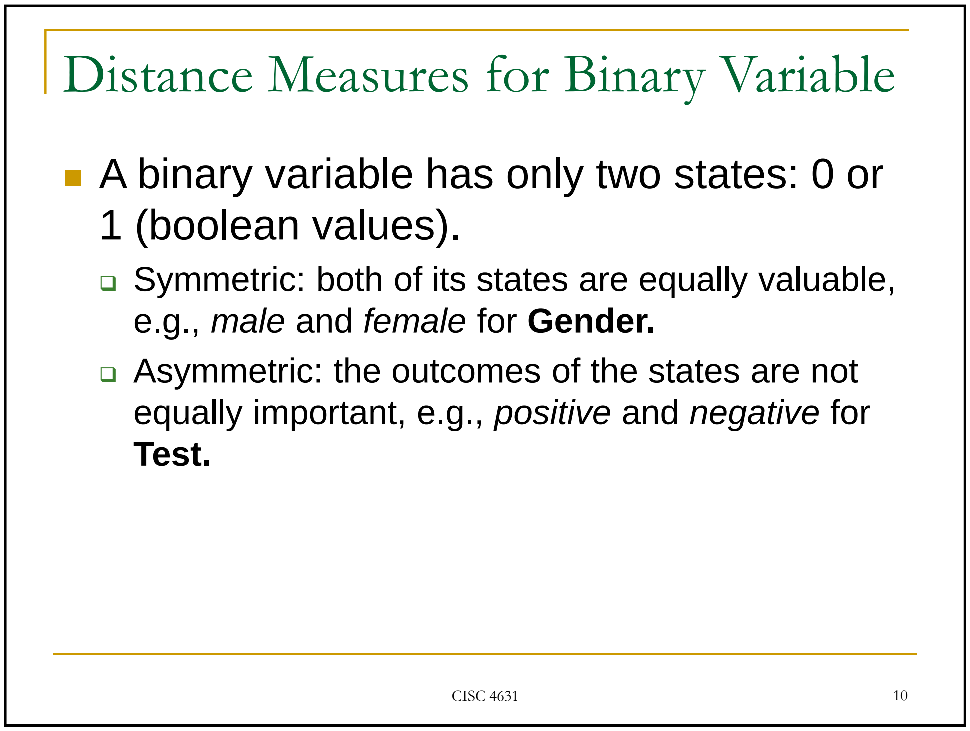
Example: Distance Measures

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

Manhattan Distance	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

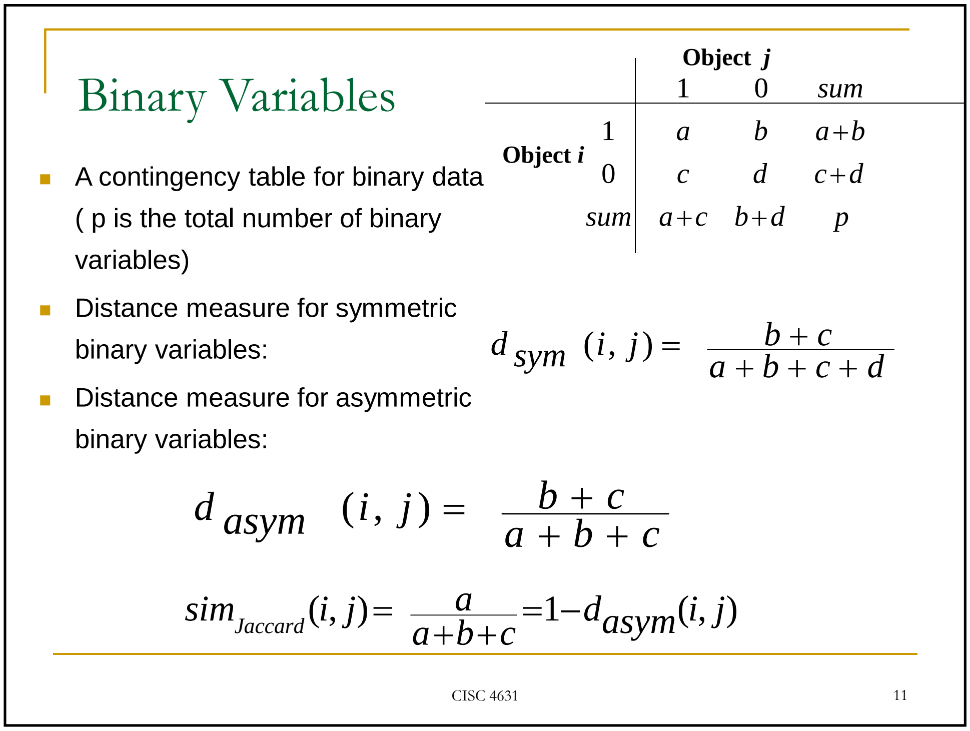
Euclidean Distance	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix



Distance Measures for Binary Variable

- A binary variable has only two states: 0 or 1 (boolean values).
  - Symmetric: both of its states are equally valuable, e.g., *male* and *female* for **Gender.**
  - Asymmetric: the outcomes of the states are not equally important, e.g., *positive* and *negative* for **Test.**



Binary Variables

- A contingency table for binary data (p is the total number of binary variables)

		Object *j*		
		1	0	*sum*
Object *i*	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Distance measure for symmetric binary variables:

$$d_{sym}(i,j) = \frac{b+c}{a+b+c+d}$$

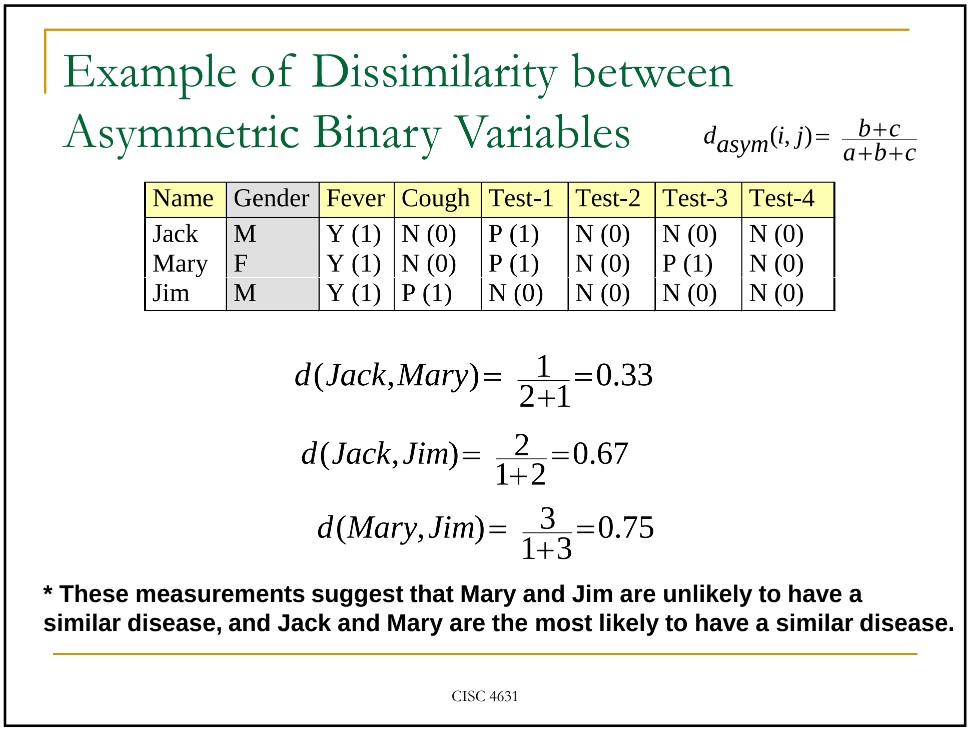
- Distance measure for asymmetric binary variables:

$$d_{asym}(i,j) = \frac{b+c}{a+b+c}$$

$$sim_{Jaccard}(i,j) = \frac{a}{a+b+c} = 1 - d_{asym}(i,j)$$

CISC 4631

11



Example of Dissimilarity between Asymmetric Binary Variables

$$d_{asym}(i,j) = \frac{b+c}{a+b+c}$$

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y (1)	N (0)	P (1)	N (0)	N (0)	N (0)
Mary	F	Y (1)	N (0)	P (1)	N (0)	P (1)	N (0)
Jim	M	Y (1)	P (1)	N (0)	N (0)	N (0)	N (0)

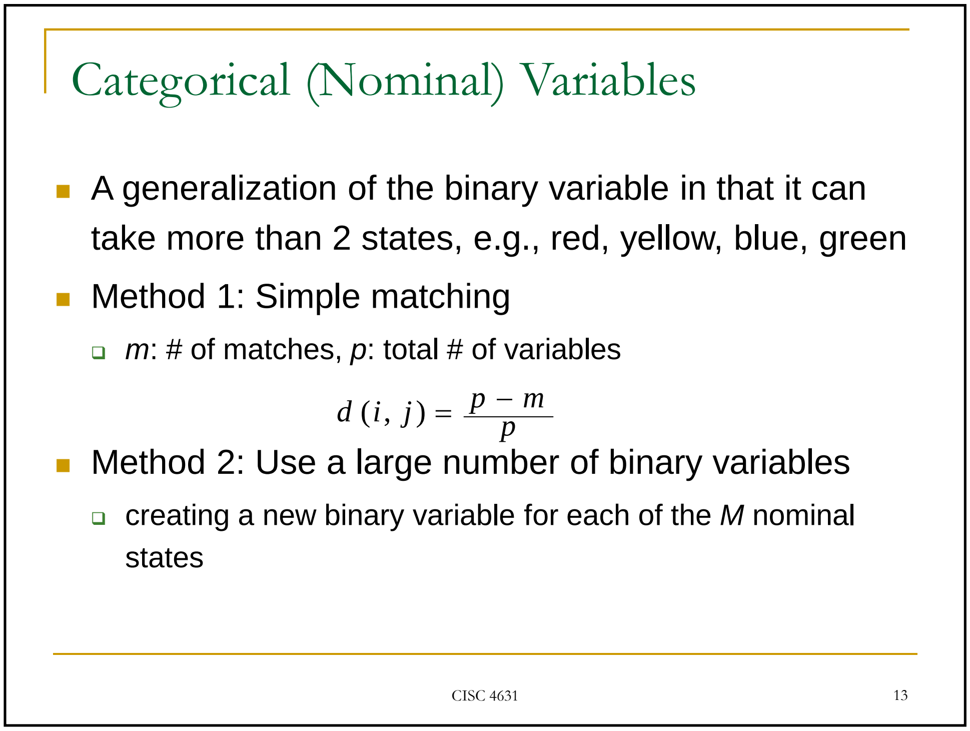
$$d(Jack, Mary) = \frac{1}{2+1} = 0.33$$

$$d(Jack, Jim) = \frac{2}{1+2} = 0.67$$

$$d(Mary, Jim) = \frac{3}{1+3} = 0.75$$

*** These measurements suggest that Mary and Jim are unlikely to have a similar disease, and Jack and Mary are the most likely to have a similar disease.**

CISC 4631

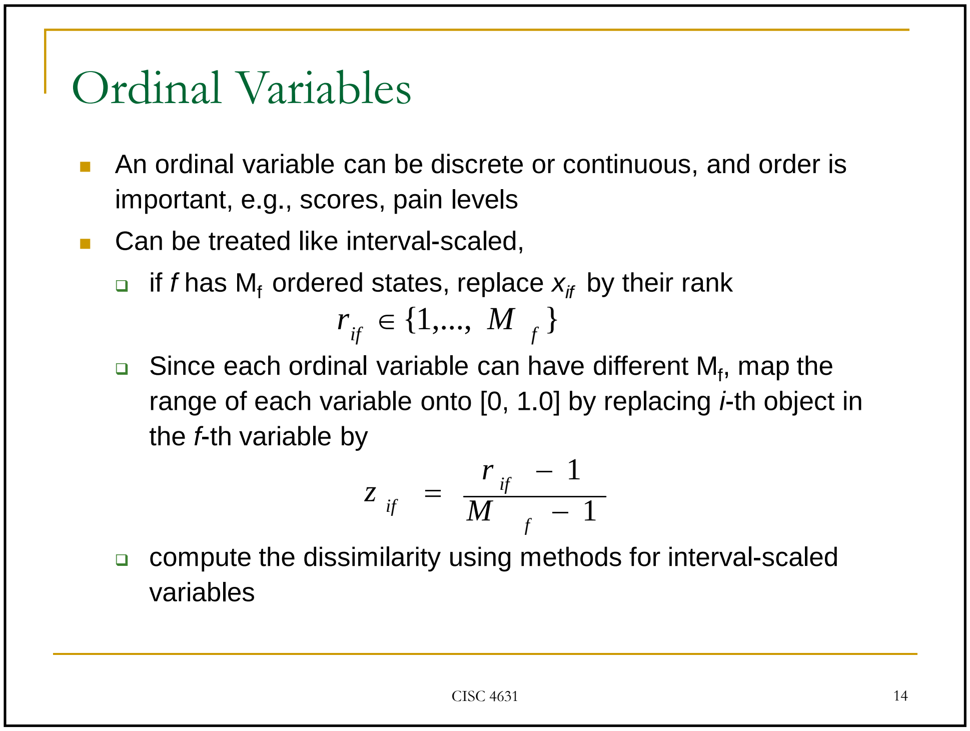


Categorical (Nominal) Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary variables
  - creating a new binary variable for each of the *M* nominal states



Ordinal Variables

- An ordinal variable can be discrete or continuous, and order is important, e.g., scores, pain levels
- Can be treated like interval-scaled,
  - if *f* has M_f ordered states, replace x_{if} by their rank

$$r_{if} \in \{1, ..., M_f\}$$

  - Since each ordinal variable can have different M_f, map the range of each variable onto [0, 1.0] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Example of Ordinal Variables

Name	Gender	Pain Levels	Blood Pressure
Jack	M	5	140/90
Mary	F	3	120/80
Jim	M	2	160/120

Pain levels (1-10):

5 -> (5-1)/(10-1) =0.44

3 -> (3-1)/(10-1) = 0.22

2 -> (2-1)/(10-1) = 0.11

Blood Pressure (High, Normal, Low):

140/90 (High - 3)->(3-1)/(3-1)=1

120/80 (Normal - 2)->(2-1)/(3-1)=0.5

160/120 (High-3) -> (3-1)/(3-1) = 1

Name	Gender	Pain Levels	Blood Pressure
Jack	M	0.44	1
Mary	F	0.22	0.5
Jim	M	0.11	1

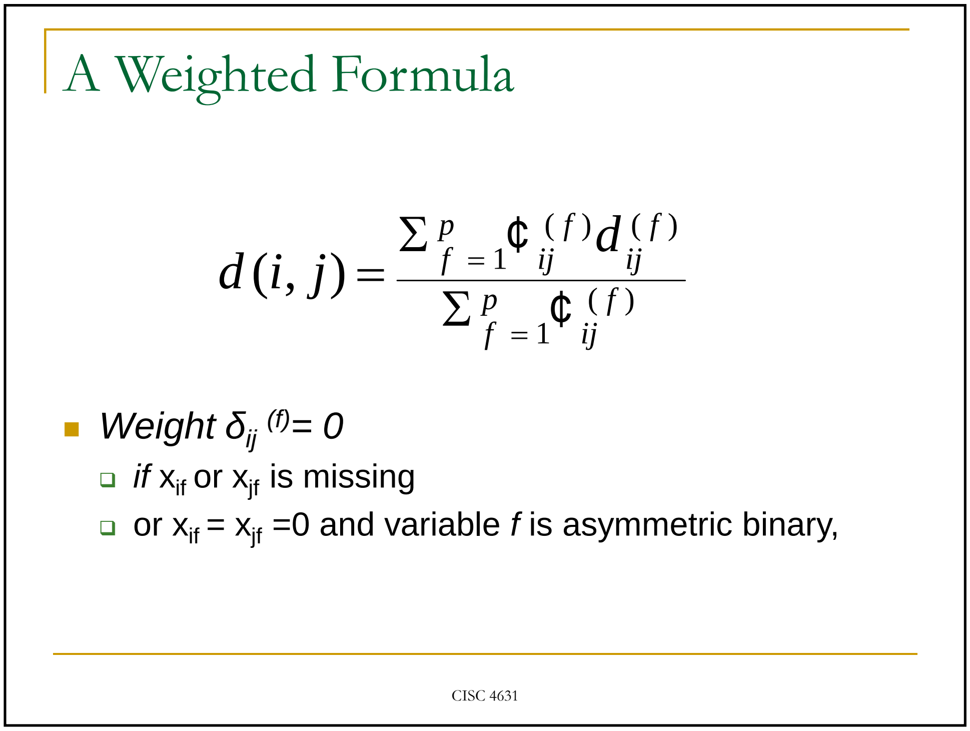
d(Jack, Mary) = $((0.44 - 0.22)^2 + (1- 0.5)^2)^{1/2}$ = 0.55

d(Jack, Jim) = $((0.44-0.11)^2 + (1-1)^2)^{1/2}$ = 0.33

d (Mary, Jim) = $((0.22-0.11)^2 + (0.5-1)^2)^{1/2}$ = 0.51

Variables of Mixed Types

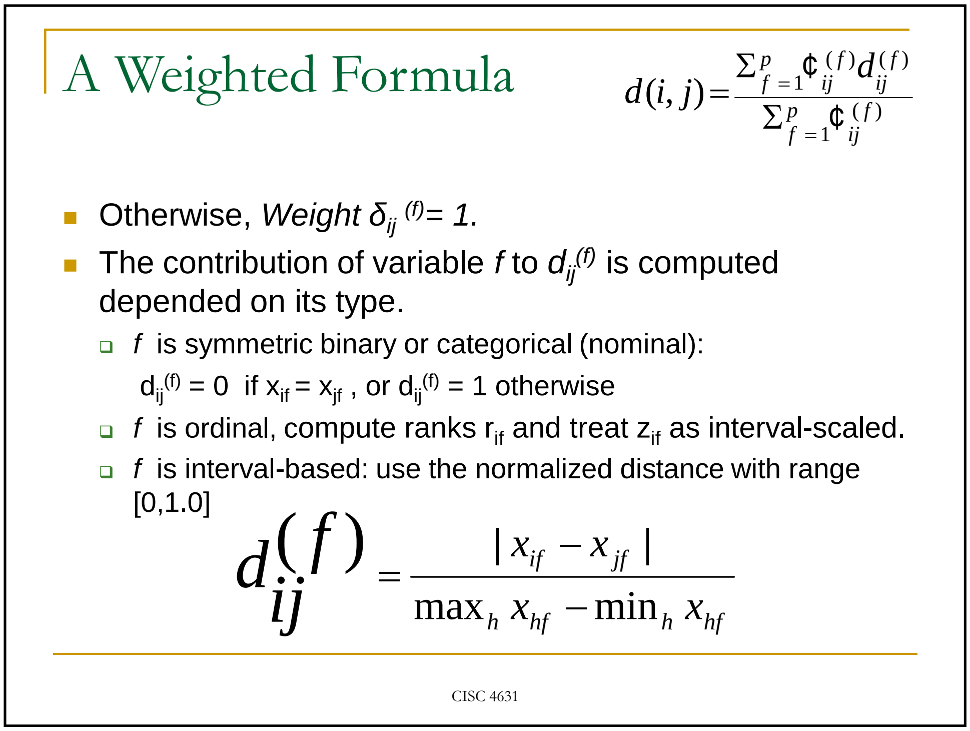
- A database may contain different types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval
- One approach is to group each type of variable together, performing a separate cluster analysis for each type.
- One approach is to bring different variables onto a common scale of the interval [0.0, 1.0], performing a single cluster analysis.
  - A weighted formula



A Weighted Formula

$$d(i,j) = \frac{\sum_{f=1}^{p} \text{¢}_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \text{¢}_{ij}^{(f)}}$$

- *Weight* $\delta_{ij}^{(f)}$ *= 0*
  - *if* x_{if} or x_{jf} is missing
  - or $x_{if} = x_{jf} = 0$ and variable *f* is asymmetric binary,



A Weighted Formula

$$d(i,j) = \frac{\sum_{f=1}^{p} \text{¢}_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \text{¢}_{ij}^{(f)}}$$

- Otherwise, *Weight* $\delta_{ij}^{(f)}$ *= 1.*
- The contribution of variable *f* to $d_{ij}^{(f)}$ is computed depended on its type.
  - *f* is symmetric binary or categorical (nominal):

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - *f* is ordinal, compute ranks r_{if} and treat z_{if} as interval-scaled.
  - *f* is interval-based: use the normalized distance with range [0,1.0]

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

Example

Name	Gender	Pain Levels	Blood Pressure	Test-1	Test-2	Test-3	Test-4
Jack	M	5	140/90	P (1)	N (0)	N (0)	N (0)
Mary	F	3	120/80	P (1)	N (0)	P (1)	N (0)
Jim	M	2	160/120	N (0)	N (0)	N (0)	N (0)

- Gender is a symmetric attribute, Pain levels and Blood pressures are ordinal, and the remaining attributes are asymmetric binary

Name	Gender	Pain Levels	Blood Pressure	Test-1	Test-2	Test-3	Test-4
Jack	M	0.44	1	P (1)	N (0)	N (0)	N (0)
Mary	F	0.22	0.5	P (1)	N (0)	P (1)	N (0)
Jim	M	0.11	1	N (0)	N (0)	N (0)	N (0)

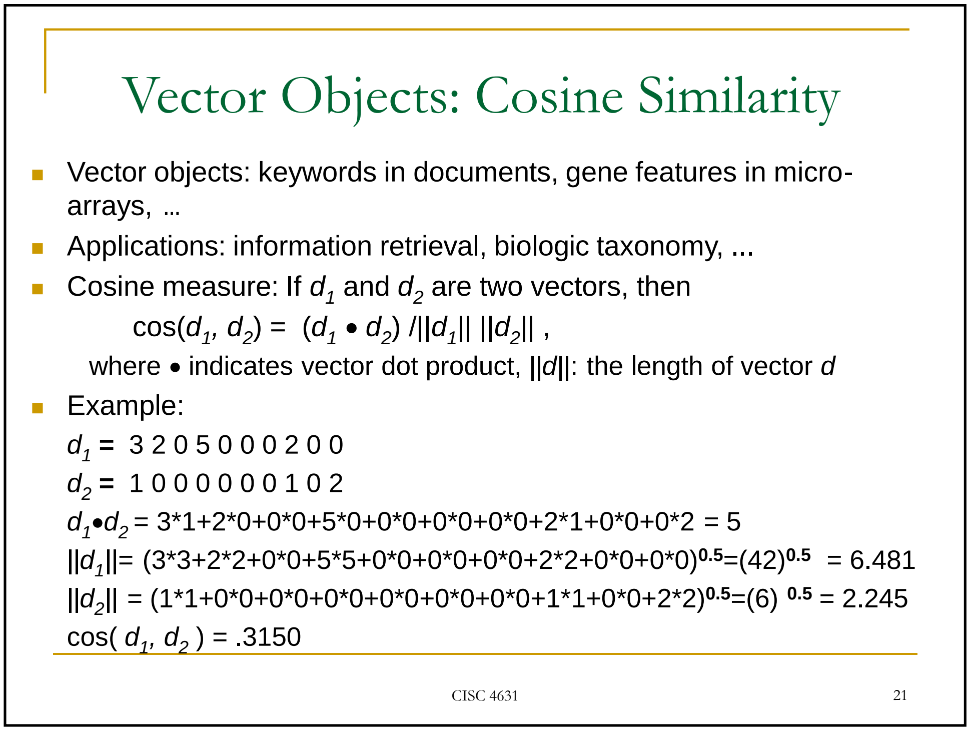
Name	Gender	Pain Levels	Blood Pressure	Test-1	Test-2	Test-3	Test-4
Jack	M	0.44	1	P (1)	N (0)	N (0)	N (0)
Mary	F	0.22	0.5	P (1)	N (0)	P (1)	N (0)
Jim	M	0.11	1	N (0)	N (0)	N (0)	N (0)

- When *i = Jack* and *j = Mary*, $\delta_{ij}^{(gender)} = 1$, $\delta_{ij}^{(Pain\ Levels)} = 1$, $\delta_{ij}^{(Blood\ Pressure)} = 1$, $\delta_{ij}^{(Test\text{-}1)} = 1$, $\delta_{ij}^{(Test\text{-}2)} = 0$, $\delta_{ij}^{(Test\text{-}3)} = 1$, $\delta_{ij}^{(Test\text{-}4)} = 0$

$$d(Jack, Mary) = \frac{1*1 + 1*\frac{|0.44-0.22|}{(0.44-0.11)} + 1*\frac{|1-0.5|}{(1-0.5)} + 1*0 + 1*1}{1+1+1+1+0+1+0} = 0.734$$

$$d(Jack, Jim) = \frac{1*0 + 1*\frac{|0.44-0.11|}{(0.44-0.11)} + 1*\frac{|1-1|}{(1-0.5)} + 1*1}{1+1+1+1+0+0+0} = 0.5$$

$$d(Jim, Mary) = \frac{1*1 + 1*\frac{|0.22-0.11|}{(0.44-0.11)} + 1*\frac{|1-0.5|}{(1-0.5)} + 1*1 + 1*1}{1+1+1+1+0+1+0} = 0.866$$



Vector Objects: Cosine Similarity

- Vector objects: keywords in documents, gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, ...
- Cosine measure: If d_1 and d_2 are two vectors, then

  $\cos(d_1, d_2) = (d_1 \bullet d_2) /||d_1|| \, ||d_2||$,

  where $\bullet$ indicates vector dot product, $||d||$: 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

  $d_1 \bullet d_2 = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5$

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

  $||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5}=(6)^{0.5} = 2.245$

  $\cos(d_1, d_2) = .3150$