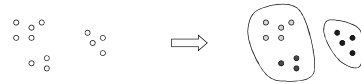


CS522 Advanced Database Systems  
Clustering: Basic Concepts and Distance Measures

Chengyu Sun  
California State University, Los Angeles

## Clustering

- ◆ Group *similar* objects together
- ◆ Applications
  - Identify users who share similar interests
  - Automatically generate concept hierarchies
  - Reduce algorithmic complexity
  - ...



## Types of Clusters

- ◆ Well separated
- ◆ Prototype based
- ◆ Contiguity based
- ◆ Density based
- ◆ Conceptual clusters

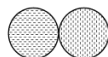
## Well-separated Clusters



- ◆ Each point is closer to all of the points in its cluster than to any point in another cluster

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## Prototype-based Clusters



- ◆ ??

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

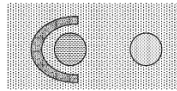
## Contiguity-based Clusters



- ◆ ??
- ◆ A cluster can be considered as a *connected component* in a graph

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

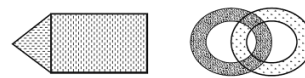
## Density-based Clusters



- ◆ A cluster is a dense region of objects surrounded by a region of low density

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## Conceptual Clusters



- ◆ A cluster is a set of objects that share *some property*

©Tan, Steinbach, Kumar Introduction to Data Mining 2004

## Type of Clustering

- |  |  |
|--|--|
| <ul style="list-style-type: none"> <li>◆ Clustering methods                     <ul style="list-style-type: none"> <li>■ Partitioning</li> <li>■ Hierarchical</li> <li>■ Density-based</li> <li>■ Grid-based</li> <li>■ Model-based</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>◆ Clustering results                     <ul style="list-style-type: none"> <li>■ Exclusive</li> <li>■ Overlapping</li> <li>■ Fuzzy</li> <li>■ Complete</li> <li>■ Partial</li> </ul> </li> </ul> |
|--|--|

## Similarity Measure

TID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No

- ◆ Is #1 more similar to #2 or #3?
- ◆ Similarity vs. Distance vs. Dissimilarity

## Interval-Scaled Attributes

- ◆ Continuous-valued data measured with a linear scale (vs. exponential or logarithmic scale)

## Distance Measures

- ◆  $\mathbf{X}=(x_1, x_2, \dots, x_n)$  and  $\mathbf{Y}=(y_1, y_2, \dots, y_n)$ 
  - E.g. (1, 2) and (3, 5)

Euclidean Distance:

$$dist(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Manhattan Distance:

$$dist(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^n |x_i - y_i|$$

## Minkowski Distance

$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p}$$

- ◆ p=1 (Manhattan Distance)
  - a.k.a. L<sub>1</sub> norm or L<sub>1</sub> distance
- ◆ p=2 (Euclidean Distance)
  - a.k.a. L<sub>2</sub> norm or L<sub>2</sub> distance

## Requirements of Distance Functions

- ◆ dist(**X**,**Y**) ≥ 0
- ◆ dist(**X**,**X**) = 0
- ◆ dist(**X**,**Y**) = dist(**Y**,**X**)
- ◆ dist(**X**,**Y**) ≤ dist(**X**,**Z**) + dist(**Z**,**Y**)
  - *Triangular Inequality*

## Problem of Units

- ◆ (10m, 2km) and (5m, 2.1km)?
- ◆ (10m, 200lb) and (5m, 210lb)?

## Standardize Interval-Scaled Attributes

- ◆ Given attribute A with values a<sub>1</sub>, a<sub>2</sub>, ..., a<sub>n</sub>

Mean:  $\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i$

Mean absolute deviation:  $s = \frac{1}{n} \sum_{i=1}^n |a_i - \bar{a}|$

Standardized measurement (*z-score*):  $z_i = \frac{a_i - \bar{a}}{s}$

## Binary Attributes

- ◆ Symmetric
  - E.g. gender
- ◆ Asymmetric
  - E.g. HIV test result

## Contingency Table for Binary Attributes

		Record <b>Y</b>	
		1	0
Record <b>X</b>	1	q	r
	0	s	t

- ◆ Example
  - X=(1,1,0,1,0,0,0), Y=(0,1,0,1,0,1,0)

## Distance Measure for Symmetric Binary Attributes

Similarity:  $sim(\mathbf{X}, \mathbf{Y}) = \frac{q+t}{q+r+s+t}$

Dissimilarity:  $dsim(\mathbf{X}, \mathbf{Y}) = \frac{r+s}{q+r+s+t}$

Distance: ??

## Distance Measure for Asymmetric Binary Attributes

Similarity (Jaccard Coefficient):  $sim(\mathbf{X}, \mathbf{Y}) = \frac{q}{q+r+s}$

Dissimilarity:  $dsim(\mathbf{X}, \mathbf{Y}) = \frac{r+s}{q+r+s}$

Distance: ??

## Binary Attribute Example

TID	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
1	M	Y	N	P	N	N	N
2	F	Y	N	P	N	P	N
3	M	Y	Y	N	N	N	N

◆  $dist(1,2)??$   $dist(2,3)??$   $dist(3,1)??$

## Categorical Attributes

### ◆ Example

- Marital status: single, married, divorced
- ◆  $dist(\mathbf{X}, \mathbf{Y}) = (p-m) / p$ 
  - $m$ : number of attribute matches
  - $p$ : total number of attributes
- ◆ Or, encode each state with a binary attribute

## Ordinal Attributes

### ◆ Example

- Grade: F, D, C, B, A
- ◆ Given an attribute with  $M$  possible values  $\{1, 2, \dots, M\}$ , map value  $a$  to the range of  $[0.0, 1.0]$

$$z = \frac{a-1}{M-1}$$

## Records with Mixed Types of Attributes ...

$$dist(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^n \delta_i \times dist(x_i, y_i)}{\sum_{i=1}^n \delta_i}$$

- ◆  $\delta_i$  is the weight of the  $i$ th attribute  $a_i$ 's contribution toward the overall distance
  - 0 if  $x_i$  or  $y_i$  is missing, or  $a_i$  is asymmetric binary and  $x_i=y_i=0$
  - 1 otherwise

## ... Records with Mixed Types of Attributes

◆  $\text{dist}(x_i, y_i)$

- Interval-based:  $|x_i - y_i| / (\max(a_i) - \min(a_i))$
- Binary or categorical: 0 if  $x_i = y_i$ ; 1 otherwise
- Ordinal: treat as interval-based using  $z_i$

## Vector Distances

Cosine distance: 
$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\|\mathbf{X}\| \|\mathbf{Y}\|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

Tanimoto distance: 
$$\text{dist}(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{X} \cdot \mathbf{Y}}{\mathbf{X} \cdot \mathbf{X} + \mathbf{Y} \cdot \mathbf{Y} - \mathbf{X} \cdot \mathbf{Y}}$$

## Readings

- ◆ Textbook 7.1, 7.2, and 7.3