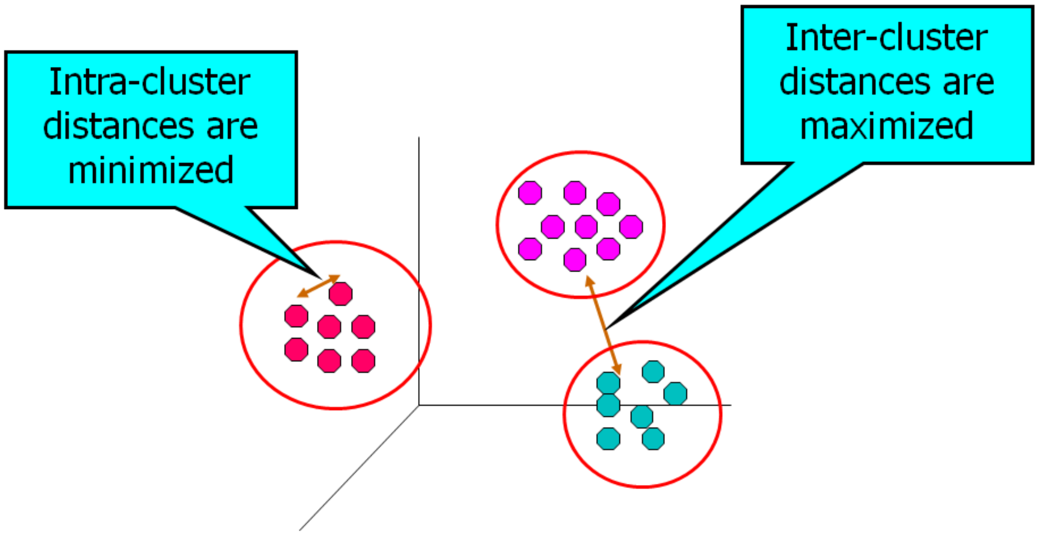
Clustering

What is Clustering?

- Process of grouping objects into classes/clusters
  - objects within a cluster are similar to one another
  - dissimilar to the objects in other clusters
- Data segmentation
  - grouping similar tuples in a database together
- Unsupervised learning: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - Pattern recognition, web search, document retrieval, business
  - As a preprocessing step for other algorithms

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity (within objects in the same cluster)
  - low inter-class similarity



Quality: What is Good Clustering?

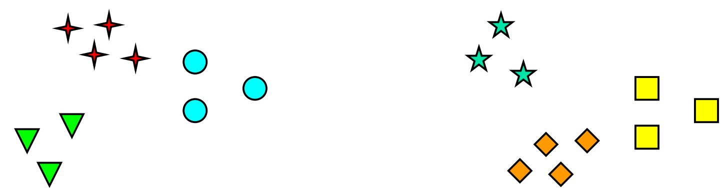
- Quality of a clustering is measured by its ability to discover some or all of the hidden patterns
- Quality of a clustering depends on the similarity measure used by the algorithm and its implementation
- Similarity/Dissimilarity metric: Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
- The definitions of distance functions are usually different for types of data
- It is hard to define "similar enough" or "good enough"
  - the answer is typically highly subjective

Conceptual Clustering

- This is different from conventional clustering
- It consists of two components
  - discover clusters
  - find descriptions for each cluster

Number of Clusters can be Ambiguous

How many clusters?



Six Clusters

Two Clusters

Four Clusters

Issues Related to Clustering in DM

- Handling outliers is difficult
- Handling noise in the data
- Interpretability
- No unique answer
- Deciding best number of clusters
- Deciding what attributes to use
- Insensitivity to the order of input data
- Dealing with different types of attributes
- Dynamic data
- Scalability
- Discovering clusters with arbitrary shapes
- Incorporation of user-specified constraints

Data Structures

- n objects, p attributes
- Data matrix
  - object-by-attribute structure
  - two-mode matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - object-by-object structure
  - one-mode matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Type of data in clustering analysis

- Interval-scaled (numerical)
- Binary
- Nominal
- Ordinal
- Variables of mixed types

Interval-Scaled Variables

- Continuous (numerical) variables
- Correspond to values of continuous measurements of a roughly linear scale
- Examples: height, temperature, income, latitude and longitude values, …
- Units used can bias performance of algorithm: meters vs. millimeters
- Values need to be standardized
  - Min-max, z-score normalization, and decimal scaling

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, ..., x_{ip})$ and $j = (x_{j1}, x_{j2}, ..., x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Similarity and Dissimilarity Between Objects (Cont.)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also, one can use weighted distance, or other dissimilarity measures

Binary Variables

- Has one of two states: 0, 1
- Examples: smoker, owns-house
- Assume that each object x_i is represented as an m-vector with each attribute value
  $x_{ik} \in \{0, 1\}$ for $1 \leq k \leq m$
- Similarity between binary variables x_i and x_j is based on using a $2\ by\ 2$ contingency table

where,

- a is number of attributes where $x_{ik} = x_{jk} = 1$
- b is number of attributes where $x_{ik} = 1$ and $x_{jk} = 0$
- c is number of attributes where $x_{ik} = 0$ and $x_{jk} = 1$
- d is number of attributes where $x_{ik} = x_{jk} = 0$

		x_j 1	x_j 0
x_i	1	a	b
	0	c	d

Binary Variables – Example

- $x_i = \{0,\ 0,\ 1,\ 1,\ 0,\ 1,\ 0,\ 1\}$
- $x_j = \{0,\ 1,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0\}$
- $a = 2$
- $b = 2$
- $c = 1$
- $d = 3$

Similarity Measures for Binary variables

- Simple matching coefficient (SMC)
- Jaccard coefficient

Simple Matching Coefficient (SMC)

- usually used for symmetric binary variables
  - both values are equally important
  - gender

		x_j	
		1	0
x_i	1	a	b
	0	c	d

$$\mathrm{sim}_{\mathrm{smc}}(\mathrm{i},\mathrm{j}) = \frac{a+d}{a+b+c+d}$$

$$\mathrm{d}_{\mathrm{smc}}(\mathrm{i},\mathrm{j}) = \frac{b+c}{a+b+c+d}$$

Jaccard Coefficient

- usually used for asymmetric binary variables
  - values are not equally important
  - HIV-positive
  - the most important value (rarest) is coded as 1

		x_j	
		1	0
x_i	1	a	b
	0	c	d

$$\mathrm{sim}_{\mathrm{Jaccard}}(\mathrm{i}, \mathrm{j}) = \frac{a}{a+b+c}$$

$$\mathrm{d}_{\mathrm{Jaccard}}(\mathrm{i}, \mathrm{j}) = \frac{b+c}{a+b+c}$$

Example

- $x_i = \{0, \ 0, \ 1, \ \ 1, \ 0, \ 1, \ \ 0, \ 1\}$
- $x_j = \{0, \ 1, \ 1, \ 0, \ 0, \ 1, \ 0, \ 0\}$
- $a = 2, b = 2, c = 1, d = 3$

$$\mathrm{d_{smc}(i, j)} = \frac{b + c}{a+b+c+d} = \frac{2+1}{2+2+1+3} = \frac{3}{8}$$

$$\mathrm{d_{Jaccard}(i, j)} = \frac{b + c}{a + b + c} = \frac{2+1}{2+2+1} = \frac{3}{5}$$

Which two objects are more similar?

- Example

Name	Gender	Smoker	Drinker	Active	Obese
Jacob	M	Y	N	Y	N
Emily	F	Y	N	Y	Y
Liam	M	Y	Y	N	N

  - gender is a symmetric attribute (not used)
  - the remaining attributes are asymmetric binary
  - let the value Y be set to 1, and the value N be set to 0

$$\mathrm{d}_{\mathrm{Jaccard}}(\mathrm{Jacob, Emily}) = \frac{b+c}{a+b+c} = \frac{0+1}{2+0+1} = \frac{1}{3}$$

$$\mathrm{d}_{\mathrm{Jaccard}}(\mathrm{Jacob, Liam}) = \frac{b+c}{a+b+c} = \frac{1+1}{1+1+1} = \frac{2}{3}$$

$$\mathrm{d}_{\mathrm{Jaccard}}(\mathrm{Emily, Liam}) = \frac{b+c}{a+b+c} = \frac{2+1}{1+2+1} = \frac{3}{4}$$

Nominal Variables

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

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

Ordinal Variables

- Similar to nominal variables but values can be ordered
- Order is important, e.g., rank
- Examples: military ranks, university professors/students
- An ordinal variable can be discrete or continuous - $\mathbb{R}$
- Can be treated like interval-scaled
  - replace x_{if} (i-th object in the f-th variable) by its rank

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

  - map the range of each variable onto [0, 1] by replacing by

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

    $x_{if} \rightarrow r_{if} \rightarrow z_{if}$

  - compute the dissimilarity using methods for interval-scaled variables

Exponential-Scaled Variables

- Aka ratio-scaled
- Take values that represent a positive measurement on a nonlinear scale, approximately at exponential scale,
  such as Ae^{Bt} or Ae^{-Bt}
- Example: decay of radioactive material, growth of bacteria
- Methods:
  - treat them like interval-scaled variables—*not a good choice!* (the scale can be distorted)
  - apply logarithmic transformation

$$y_{if} = log(x_{if})$$

  - treat them as continuous ordinal variables and treat their ranks as interval-scaled values

Variables of Mixed Types

- A database may contain different types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio
- One may use a single formula to compute similarity

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

  - where indicator,

$\delta_{ij}(f) = 0$ if either 1) value is missing or
2) f is binary asymmetric, and
value is 0 under both vectors

$\delta_{ij}(f) = 1$ otherwise

Variables of Mixed Types

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

- p is number of attributes
- $d_{ij}(f)$ is the contribution of attribute f to the similarity and is based on its type
- if f is binary or nominal:
  $d_{ij}(f)$ = 0 if x_{if} = x_{jf}
  $d_{ij}(f)$ = 1 otherwise
- if f is ordinal or ratio-scaled:
  compute the ranks, r_{if}, and
  treat z_{if} as interval-scaled

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

- If f is numerical, then $d_{ij}(f) = \frac{|x_{if} - x_{jf}|}{\max\{x_f\} - \min\{x_f\}}$

Cosine Similarity

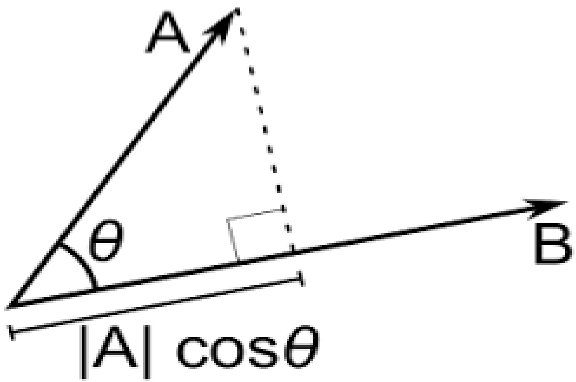
- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular term in the document
- Each document is represented as a term-frequency vector

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: bioinformatics data
- Applications: information retrieval, text data mining, bioinformatics

Cosine Similarity

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*
Document1	5	0	3	0	2	0
Document2	3	0	2	0	1	1
Document3	0	7	0	2	1	0
Document4	0	1	0	0	1	2



- A similarity measure for vector data needs to ignore 0 matches and be able to handle non-binary data
- Cosine measure: If $\boldsymbol{d_1}$ and $\boldsymbol{d_2}$ are two vectors, then

$$sim(\boldsymbol{d_1}, \boldsymbol{d}_2) = cos(\boldsymbol{d_1}, \boldsymbol{d}_2) = \frac{\boldsymbol{d_1} \bullet \boldsymbol{d_2}}{\| \boldsymbol{d_1} \| \| \boldsymbol{d_2} \|}$$

where $\bullet$ indicates vector dot product, and $\| \boldsymbol{d} \|$ is the Euclidean length of the vector $\boldsymbol{d}$

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$,
  where $\bullet$ indicates vector dot product, $||d||$: the length of vector d

- Ex: Find the **similarity** between documents 1 and 2.

$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{\mathbf{0.5}} = (42)^{\mathbf{0.5}}$
$= 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{\mathbf{0.5}} = (17)^{\mathbf{0.5}}$
$= 4.12$
$\cos(d_1, d_2) = 25 / (6.481 * 4.12) = 0.94$