Clustering
Lecture 1: Basics

Jing Gao
SUNY Buffalo
Class Structure

• **Topics**
  – Clustering, Classification
  – Network mining
  – Anomaly detection

• **Expectation**
  – Sign-in
  – Take quiz in class
  – Two more projects on clustering and classification
  – One more homework on network mining or anomaly detection

• **Website**
  – http://www.cse.buffalo.edu/~jing/cse601/fa12/
Outline

• **Basics**
  – Motivation, definition, evaluation

• **Methods**
  – Partitional,
  – Hierarchical
  – Density-based
  – Mixture model
  – Spectral methods

• **Advanced topics**
  – Clustering ensemble
  – Clustering in MapReduce
  – Semi-supervised clustering, subspace clustering, co-clustering, etc.
Readings

• Tan, Steinbach, Kumar, Chapters 8 and 9.
• Han, Kamber, Pei. Data Mining: Concepts and Techniques. Chapters 10 and 11.
• Additional readings posted on website
Clustering Basics

• Definition and Motivation
• Data Preprocessing and Similarity Computation
• Objective of Clustering
• Clustering Evaluation
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.
Application Examples

• A stand-alone tool: explore data distribution
• A preprocessing step for other algorithms
• Pattern recognition, spatial data analysis, image processing, market research, WWW, ...
  – Cluster documents
  – Cluster web log data to discover groups of similar access patterns
Clustering Co-expressed Genes

Gene Expression Data Matrix

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.3956</td>
<td>-1.1906</td>
<td>-1.6109</td>
<td>-0.0176</td>
<td>-0.4958</td>
<td>-0.3956</td>
<td>-0.59544</td>
</tr>
<tr>
<td>2</td>
<td>-1.2081</td>
<td>-0.1064</td>
<td>-1.4051</td>
<td>0.408471</td>
<td>-0.5248</td>
<td>-0.40845</td>
<td>0.66481</td>
</tr>
<tr>
<td>3</td>
<td>0.66037</td>
<td>-0.59889</td>
<td>-1.87056</td>
<td>-1.76302</td>
<td>-0.9852</td>
<td>-0.28403</td>
<td>-0.37445</td>
</tr>
<tr>
<td>4</td>
<td>0.31654</td>
<td>-1.50006</td>
<td>-1.30162</td>
<td>-1.59723</td>
<td>-0.54944</td>
<td>0.24360</td>
<td>0.90647</td>
</tr>
<tr>
<td>5</td>
<td>0.309699</td>
<td>0.07136</td>
<td>-0.9559</td>
<td>0.312404</td>
<td>-0.20826</td>
<td>-0.25584</td>
<td>-1.7902</td>
</tr>
<tr>
<td>6</td>
<td>-1.42309</td>
<td>0.01419</td>
<td>-1.10109</td>
<td>-0.20726</td>
<td>-0.42089</td>
<td>-0.9167</td>
<td>-0.00517</td>
</tr>
<tr>
<td>7</td>
<td>0.63939</td>
<td>1.20096</td>
<td>-1.29059</td>
<td>0.359066</td>
<td>-0.5467</td>
<td>0.60529</td>
<td>-0.3195</td>
</tr>
<tr>
<td>8</td>
<td>0.70912</td>
<td>-0.71537</td>
<td>-1.08584</td>
<td>-0.00574</td>
<td>-0.52383</td>
<td>-0.37029</td>
<td>-0.42753</td>
</tr>
<tr>
<td>9</td>
<td>0.34727</td>
<td>-1.02277</td>
<td>-1.01239</td>
<td>-0.33022</td>
<td>0.33033</td>
<td>0.602159</td>
<td>0.919311</td>
</tr>
<tr>
<td>10</td>
<td>-1.72998</td>
<td>-0.04202</td>
<td>-1.04118</td>
<td>0.13236</td>
<td>-0.16425</td>
<td>-0.25339</td>
<td>0.522359</td>
</tr>
<tr>
<td>11</td>
<td>-1.54734</td>
<td>-0.59666</td>
<td>-1.32333</td>
<td>0.011196</td>
<td>-0.36393</td>
<td>0.0496</td>
<td>0.342714</td>
</tr>
<tr>
<td>12</td>
<td>0.734183</td>
<td>0.36694</td>
<td>-1.74561</td>
<td>0.179796</td>
<td>-0.38027</td>
<td>0.16517</td>
<td>-0.94872</td>
</tr>
<tr>
<td>13</td>
<td>0.02909</td>
<td>-1.12806</td>
<td>-1.65255</td>
<td>0.3803</td>
<td>0.095634</td>
<td>0.00496</td>
<td>-0.63838</td>
</tr>
<tr>
<td>14</td>
<td>-0.75279</td>
<td>-0.88276</td>
<td>-1.20572</td>
<td>-0.73035</td>
<td>-0.79123</td>
<td>0.14737</td>
<td>0.257327</td>
</tr>
<tr>
<td>15</td>
<td>0.011</td>
<td>0.120063</td>
<td>-1.162911</td>
<td>-0.16028</td>
<td>-1.10363</td>
<td>-1.23861</td>
<td>-0.42942</td>
</tr>
</tbody>
</table>

Gene Expression Patterns

Why looking for co-expressed genes?

- Co-expression indicates co-function;
- Co-expression also indicates co-regulation.

Co-expressed Genes
Gene-based Clustering

Examples of co-expressed genes and coherent patterns in gene expression data

Iyer’s data [2]

Other Applications

• Land use: Identification of areas of similar land use in an earth observation database
• Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
• City-planning: Identifying groups of houses according to their house type, value, and geographical location
• Climate: understanding earth climate, find patterns of atmospheric and ocean
Two Important Aspects

• **Properties of input data**
  – Define the similarity or dissimilarity between points

• **Requirement of clustering**
  – Define the objective and methodology
Clustering Basics

• Definition and Motivation
• Data Preprocessing and Distance computation
• Objective of Clustering
• Clustering Evaluation
Data Representation

- Data: Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as dimension, variable, field, characteristic, or feature

- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Data Matrix

• Represents \( n \) objects with \( p \) variables
  – An \( n \) by \( p \) matrix

\[
\begin{bmatrix}
  x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nf} & \cdots & x_{np}
\end{bmatrix}
\]

The value of the \( i \)-th object on the \( f \)-th attribute
<table>
<thead>
<tr>
<th>gene</th>
<th>sample 1</th>
<th>sample 2</th>
<th>sample 3</th>
<th>sample 4</th>
<th>sample ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>0.13</td>
<td>0.72</td>
<td>0.1</td>
<td>0.57</td>
<td>...</td>
</tr>
<tr>
<td>gene 2</td>
<td>0.34</td>
<td>1.58</td>
<td>1.05</td>
<td>1.15</td>
<td>...</td>
</tr>
<tr>
<td>gene 3</td>
<td>0.43</td>
<td>1.1</td>
<td>0.97</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>gene 4</td>
<td>1.22</td>
<td>0.97</td>
<td>1</td>
<td>0.85</td>
<td>...</td>
</tr>
<tr>
<td>gene 5</td>
<td>-0.89</td>
<td>1.21</td>
<td>1.29</td>
<td>1.08</td>
<td>...</td>
</tr>
<tr>
<td>gene 6</td>
<td>1.1</td>
<td>1.45</td>
<td>1.44</td>
<td>1.12</td>
<td>...</td>
</tr>
<tr>
<td>gene 7</td>
<td>0.83</td>
<td>1.15</td>
<td>1.1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>gene 8</td>
<td>0.87</td>
<td>1.32</td>
<td>1.35</td>
<td>1.13</td>
<td>...</td>
</tr>
<tr>
<td>gene 9</td>
<td>-0.33</td>
<td>1.01</td>
<td>1.38</td>
<td>1.21</td>
<td>...</td>
</tr>
<tr>
<td>gene 10</td>
<td>0.10</td>
<td>0.85</td>
<td>1.03</td>
<td>1</td>
<td>...</td>
</tr>
</tbody>
</table>

- **Clustering genes**
  - Genes are objects
  - Experiment conditions are attributes
  - Find genes with similar function
Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Is higher when objects are more alike
  - Often falls in the range \([0,1]\)

- **Dissimilarity**
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
Distance Matrix

- Represents pairwise distance in $n$ objects
  - An $n$ by $n$ matrix
  - $d(i,j)$: distance or dissimilarity between objects $i$ and $j$
  - Nonnegative
  - Close to 0: similar

\[
\begin{bmatrix}
0 & d(2,1) & 0 \\
\vdots & \vdots & \vdots \\
d(n,1) & d(n,2) & \cdots & \cdots & 0
\end{bmatrix}
\]
### Data Matrix -> Distance Matrix

<table>
<thead>
<tr>
<th></th>
<th>s 1</th>
<th>s 2</th>
<th>s 3</th>
<th>s 4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>g 1</td>
<td>0.13</td>
<td>0.72</td>
<td>0.1</td>
<td>0.57</td>
<td></td>
</tr>
<tr>
<td>g 2</td>
<td>0.34</td>
<td>1.58</td>
<td>1.05</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>g 3</td>
<td>0.43</td>
<td>1.1</td>
<td>0.97</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>g 4</td>
<td>1.22</td>
<td>0.97</td>
<td>1</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>g 5</td>
<td>-0.89</td>
<td>1.21</td>
<td>1.29</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>g 6</td>
<td>1.1</td>
<td>1.45</td>
<td>1.44</td>
<td>1.12</td>
<td></td>
</tr>
<tr>
<td>g 7</td>
<td>0.83</td>
<td>1.15</td>
<td>1.1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>g 8</td>
<td>0.87</td>
<td>1.32</td>
<td>1.35</td>
<td>1.13</td>
<td></td>
</tr>
<tr>
<td>g 9</td>
<td>-0.33</td>
<td>1.01</td>
<td>1.38</td>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td>g 10</td>
<td>0.10</td>
<td>0.85</td>
<td>1.03</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**Original Data Matrix**

<table>
<thead>
<tr>
<th></th>
<th>g 1</th>
<th>g 2</th>
<th>g 3</th>
<th>g 4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>g 1</td>
<td>0</td>
<td>d(1,2)</td>
<td>d(1,3)</td>
<td>d(1,4)</td>
<td></td>
</tr>
<tr>
<td>g 2</td>
<td>0</td>
<td>0</td>
<td>d(2,3)</td>
<td>d(2,4)</td>
<td></td>
</tr>
<tr>
<td>g 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>d(3,4)</td>
<td></td>
</tr>
<tr>
<td>g 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Distance Matrix**
### Types of Attributes

**Discrete**
- Has only a finite or countably infinite set of values
- Examples: zip codes, counts, or the set of words in a collection of documents
- Note: binary attributes are a special case of discrete attributes

**Ordinal**
- Has only a finite or countably infinite set of values
- Order of values is important
- Examples: rankings (e.g., pain level 1-10), grades (A, B, C, D)

**Continuous**
- Has real numbers as attribute values
- Examples: temperature, height, or weight
- Continuous attributes are typically represented as floating-point variables
## Similarity/Dissimilarity for Simple Attributes

$p$ and $q$ are the attribute values for two data objects.

<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Dissimilarity</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discrete</strong></td>
<td>$d = \begin{cases} 0 &amp; \text{if } p = q \ 1 &amp; \text{if } p \neq q \end{cases}$</td>
<td>$s = \begin{cases} 1 &amp; \text{if } p = q \ 0 &amp; \text{if } p \neq q \end{cases}$</td>
</tr>
<tr>
<td><strong>Ordinal</strong></td>
<td>$d = \frac{</td>
<td>p-q</td>
</tr>
<tr>
<td><strong>Continuous</strong></td>
<td>$d =</td>
<td>p - q</td>
</tr>
</tbody>
</table>

Dissimilarity and similarity between $p$ and $q$
Minkowski Distance—Continuous Attribute

- Minkowski distance: a generalization

\[ d(i, j) = \sqrt[q]{\sum_{p=1}^{q} |x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q} \quad (q > 0) \]

- If \( q = 2 \), \( d \) is Euclidean distance
- If \( q = 1 \), \( d \) is Manhattan distance
Standardization

• Calculate the mean absolute deviation

\[ m_f = \frac{1}{n} (x_{1f} + x_{2f} + \ldots + x_{nf}). \]

\[ s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|) \]

• Calculate the standardized measurement (z-score)

\[ z_{if} = \frac{x_{if} - m_f}{s_f} \]
Mahalanobis Distance

\[ d(p, q) = (p - q) \sum^{-1} (p - q)^T \]

\( \sum \) is the covariance matrix of the input data \( X \)

\[ \Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k) \]

Belongs to the family of bregman divergence

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.
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

1. \( d(p, q) \geq 0 \) for all \( p \) and \( q \) and \( d(p, q) = 0 \) only if \( p = q \). (Positive definiteness)
2. \( d(p, q) = d(q, p) \) for all \( p \) and \( q \). (Symmetry)
3. \( d(p, r) \leq d(p, q) + d(q, r) \) for all points \( p, q, \) and \( r \). (Triangle Inequality)

where \( d(p, q) \) is the distance (dissimilarity) between points (data objects), \( p \) and \( q \).

- A distance that satisfies these properties is a metric
Similarity for Binary Attributes

• Common situation is that objects, \( p \) and \( q \), have only binary attributes

• Compute similarities using the following quantities
  
  \[ M_{01} = \text{the number of attributes where } p \text{ was 0 and } q \text{ was 1} \]
  \[ M_{10} = \text{the number of attributes where } p \text{ was 1 and } q \text{ was 0} \]
  \[ M_{00} = \text{the number of attributes where } p \text{ was 0 and } q \text{ was 0} \]
  \[ M_{11} = \text{the number of attributes where } p \text{ was 1 and } q \text{ was 1} \]

• Simple Matching and Jaccard Coefficients

  \[ \text{SMC} = \frac{\text{number of matches}}{\text{total number of attributes}} \]
  \[ = \frac{(M_{11} + M_{00})}{(M_{01} + M_{10} + M_{11} + M_{00})} \]

  \( J \) = number of matches / number of not-both-zero attributes values
  \[ = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \]
SMC versus Jaccard: Example

\[ p = 10000000000 \]
\[ q = 0000001001 \]

\[ M_{01} = 2 \quad (\text{the number of attributes where } p \text{ was 0 and } q \text{ was 1}) \]
\[ M_{10} = 1 \quad (\text{the number of attributes where } p \text{ was 1 and } q \text{ was 0}) \]
\[ M_{00} = 7 \quad (\text{the number of attributes where } p \text{ was 0 and } q \text{ was 0}) \]
\[ M_{11} = 0 \quad (\text{the number of attributes where } p \text{ was 1 and } q \text{ was 1}) \]

\[ \text{SMC} = \frac{(M_{11} + M_{00})}{(M_{01} + M_{10} + M_{11} + M_{00})} = \frac{0+7}{2+1+0+7} = 0.7 \]

\[ J = \frac{(M_{11})}{(M_{01} + M_{10} + M_{11})} = 0 / (2 + 1 + 0) = 0 \]
Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>play</th>
<th>ball</th>
<th>score</th>
<th>game</th>
<th>win</th>
<th>lost</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Document 2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>
Cosine Similarity

• If \( d_1 \) and \( d_2 \) are two document vectors, then
  \[
  \cos( d_1, d_2 ) = \frac{d_1 \cdot d_2}{||d_1|| \ ||d_2||},
  \]
  where \( \cdot \) indicates vector dot product and \( ||d|| \) is the length of vector \( d \).

• Example:

\[
\begin{align*}
  d_1 &= 3205000200 \\
  d_2 &= 1000000102
\end{align*}
\]

\[
\begin{align*}
  d_1 \cdot d_2 &= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5 \\
  ||d_1|| &= (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481 \\
  ||d_2|| &= (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245 \\
  \cos( d_1, d_2 ) &= .3150
\end{align*}
\]
Correlation

• Correlation measures the linear relationship between objects
• To compute correlation, we standardize data objects, p and q, and then take their dot product (continuous attributes)

\[
p'_k = (p_k - \text{mean}(p)) / \text{std}(p) \\
q'_k = (q_k - \text{mean}(q)) / \text{std}(q) \\
s(p, q) = p' \cdot q'
\]
Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all $p$ and $q$. (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), $p$ and $q$. 
Characteristics of the Input Data Are Important

- Sparseness
- Attribute type
- Type of Data
- Dimensionality
- Noise and Outliers
- Type of Distribution

=> Conduct preprocessing and select the appropriate dissimilarity or similarity measure

=> Determine the objective of clustering and choose the appropriate method
Clustering Basics

• Definition and Motivation
• Data Preprocessing and Distance computation
• Objective of Clustering
• Clustering Evaluation
Considerations for Cluster Analysis

• **Partitioning criteria**
  – Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)

• **Separation of clusters**
  – Exclusive (e.g., one customer belongs to only one region) vs. overlapping (e.g., one document may belong to more than one topic)

• **Hard versus fuzzy**
  – In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  – Weights must sum to 1
  – Probabilistic clustering has similar characteristics

• **Similarity measure and data types**

• **Heterogeneous versus homogeneous**
  – Cluster of widely different sizes, shapes, and densities
Requirements of Clustering

• Scalability
• Ability to deal with different types of attributes
• Minimal requirements for domain knowledge to determine input parameters
• Able to deal with noise and outliers
• Discovery of clusters with arbitrary shape
• Insensitive to order of input records
• High dimensionality
• Incorporation of user-specified constraints
• Interpretability and usability

• What clustering results we want to get?
Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters
Partitional Clustering

Input Data

A Partitional Clustering
Hierarchical Clustering

Clustering Solution 1

Clustering Solution 2
Types of Clusters: Center-Based

• **Center-based**
  – A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
  – The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster

4 center-based clusters
Types of Clusters: Density-Based

• **Density-based**
  – A cluster is a *dense region of points*, which is *separated by low-density regions*, from other regions of high density.
  – Used when the clusters are irregular or intertwined, and when noise and outliers are present.

6 density-based clusters
Clustering Basics

• Definition and Motivation
• Data Preprocessing and Distance computation
• Objective of Clustering
• Clustering Evaluation
Cluster Validation

• **Cluster validation**
  – Quality: “goodness” of clusters
  – Assess the quality and reliability of clustering results

• **Why validation?**
  – To avoid finding clusters formed by chance
  – To compare clustering algorithms
  – To choose clustering parameters
    • e.g., the number of clusters
Aspects of Cluster Validation

• Comparing the clustering results to *ground truth* (externally known results)
  – External Index
• Evaluating the quality of clusters *without* reference to external information
  – Use only the data
  – Internal Index
• Determining the *reliability* of clusters
  – To what confidence level, the clusters are not formed by chance
  – Statistical framework
Comparing to Ground Truth

• **Notation**
  – $N$: number of objects in the data set
  – $P = \{P_1, \ldots, P_s\}$: the set of “ground truth” clusters
  – $C = \{C_1, \ldots, C_t\}$: the set of clusters reported by a clustering algorithm

• **The “incidence matrix”**
  – $N \times N$ (both rows and columns correspond to objects)
  – $P_{ij} = 1$ if $O_i$ and $O_j$ belong to the same “ground truth” cluster in $P$; $P_{ij} = 0$ otherwise
  – $C_{ij} = 1$ if $O_i$ and $O_j$ belong to the same cluster in $C$; $C_{ij} = 0$ otherwise
Rand Index and Jaccard Coefficient

A pair of data object \((O_i, O_j)\) falls into one of the following categories

- **SS**: \(C_{ij} = 1\) and \(P_{ij} = 1\); (agree)
- **DD**: \(C_{ij} = 0\) and \(P_{ij} = 0\); (agree)
- **SD**: \(C_{ij} = 1\) and \(P_{ij} = 0\); (disagree)
- **DS**: \(C_{ij} = 0\) and \(P_{ij} = 1\); (disagree)

**Rand index**

\[
\text{Rand} = \frac{|\text{Agree}|}{|\text{Agree}| + |\text{Disagree}|} = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|}
\]

- may be dominated by DD

**Jaccard Coefficient**

\[
\text{Jaccard coefficient} = \frac{|SS|}{|SS| + |SD| + |DS|}
\]
### Clustering

<table>
<thead>
<tr>
<th></th>
<th>g1</th>
<th>g2</th>
<th>g3</th>
<th>g4</th>
<th>g5</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Groundtruth

<table>
<thead>
<tr>
<th></th>
<th>g1</th>
<th>g2</th>
<th>g3</th>
<th>g4</th>
<th>g5</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>g3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### Rand

\[
\text{Rand} = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|} = \frac{17}{25}
\]

### Jaccard

\[
\text{Jaccard} = \frac{|SS|}{|SS| + |SD| + |DS|} = \frac{9}{17}
\]
Entropy and Purity

• **Notation**
  
  - $|C_k \cap P_j|$ the number of objects in both the $k$-th cluster of the clustering solution and $j$-th cluster of the groundtruth
  
  - $|C_k|$ the number of objects in the $k$-th cluster of the clustering solution
  
  - $|P_j|$ the number of objects in the $j$-th cluster of the groundtruth

• **Purity**
  
  $Purity = \frac{1}{N} \sum_k \max_j |C_k \cap P_j|$

• **Normalized Mutual Information**
  
  $NMI = \frac{I(C,P)}{\sqrt{H(C)H(P)}}$

  \[ I(C,P) = \sum_k \sum_j \frac{|C_k \cap P_j|}{N} \log \frac{N \cdot |C_k \cap P_j|}{|C_k| \cdot |P_j|} \]

  \[ H(C) = \sum_k \frac{|C_k|}{N} \log \frac{|C_k|}{N} \]

  \[ H(P) = \sum_j \frac{|P_j|}{N} \log \frac{|P_j|}{N} \]
Example

<table>
<thead>
<tr>
<th></th>
<th>P 1</th>
<th>P 2</th>
<th>P 3</th>
<th>P 4</th>
<th>P5</th>
<th>P6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>3</td>
<td>5</td>
<td>40</td>
<td>506</td>
<td>96</td>
<td>27</td>
<td>677</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>7</td>
<td>280</td>
<td>29</td>
<td>39</td>
<td>2</td>
<td>361</td>
</tr>
<tr>
<td>C3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>671</td>
<td>685</td>
</tr>
<tr>
<td>C4</td>
<td>10</td>
<td>162</td>
<td>3</td>
<td>119</td>
<td>73</td>
<td>2</td>
<td>369</td>
</tr>
<tr>
<td>C5</td>
<td>331</td>
<td>22</td>
<td>5</td>
<td>70</td>
<td>13</td>
<td>23</td>
<td>464</td>
</tr>
<tr>
<td>C6</td>
<td>5</td>
<td>358</td>
<td>12</td>
<td>212</td>
<td>48</td>
<td>13</td>
<td>648</td>
</tr>
<tr>
<td>total</td>
<td>354</td>
<td>555</td>
<td>341</td>
<td>943</td>
<td>273</td>
<td>738</td>
<td>3204</td>
</tr>
</tbody>
</table>

\[ Purity = \frac{1}{N} \sum_{k} \max_{j} |C_k \cap P_j| \]

\[ Purity = \frac{506 + 280 + 671 + 162 + 331 + 358}{3204} = 0.7203 \]

\[ NMI = \frac{I(C, P)}{\sqrt{H(C)H(P)}} \]

\[ I(C, P) = \sum_{k} \sum_{j} \frac{|C_k \cap P_j|}{N} \log \frac{N \cdot |C_k \cap P_j|}{|C_k \cap P_j|} \]

\[ H(C) = \sum_{k} \frac{|C_k|}{N} \log \frac{|C_k|}{N} \]

\[ H(P) = \sum_{j} \frac{|P_j|}{N} \log \frac{|P_j|}{N} \]
Internal Index

- “Ground truth” may be unavailable
- Use only the data to measure cluster quality
  - Measure the “cohesion” and “separation” of clusters
  - Calculate the correlation between clustering results and distance matrix
Cohesion and Separation

- **Cohesion** is measured by the within cluster sum of squares
  
  \[ WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2 \]

- **Separation** is measured by the between cluster sum of squares
  
  \[ BSS = \sum_{i} |C_i| (m - m_i)^2 \]
  
  where \(|C_i|\) is the size of cluster \(i\), \(m\) is the centroid of the whole data set

- \(BSS + WSS = \text{constant}\)
- \(WSS\) (Cohesion) measure is called Sum of Squared Error (SSE)—a commonly used measure
- A larger number of clusters tend to result in smaller SSE
Example

\[ WSS = (1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10 \]

**K=1:**

\[ BSS = 4 \times (3 - 3)^2 = 0 \]

Total = 10 + 0 = 10

**K=2:**

\[ WSS = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1 \]

\[ BSS = 2 \times (3 - 1.5)^2 + 2 \times (4.5 - 3)^2 = 9 \]

Total = 1 + 9 = 10

**K=4:**

\[ WSS = (1 - 1)^2 + (2 - 2)^2 + (4 - 4)^2 + (5 - 5)^2 = 0 \]

\[ BSS = 1 \times (1 - 3)^2 + 1 \times (2 - 3)^2 + 1 \times (4 - 3)^2 + 1 \times (5 - 3)^2 = 10 \]

Total = 0 + 10 = 10
Silhouette Coefficient

- Silhouette Coefficient combines ideas of both cohesion and separation

- For an individual point, $i$
  - Calculate $a =$ average distance of $i$ to the points in its cluster
  - Calculate $b =$ min (average distance of $i$ to points in another cluster)
  - The **silhouette coefficient** for a point is then given by
    
    $s = 1 - \frac{a}{b}$ if $a < b$, \hspace{1em} \( s = \frac{b}{a} - 1 \) if $a \geq b$, not the usual case

  - Typically between 0 and 1
  - The closer to 1 the better

- Can calculate the Average Silhouette width for a cluster or a clustering
Correlation with Distance Matrix

• Distance Matrix
  – \( D_{ij} \) is the similarity between object \( O_i \) and \( O_j \)

• Incidence Matrix
  – \( C_{ij} = 1 \) if \( O_i \) and \( O_j \) belong to the same cluster, \( C_{ij} = 0 \) otherwise

• Compute the correlation between the two matrices
  – Only \( n(n-1)/2 \) entries needs to be calculated

• High correlation indicates good clustering
Given Distance Matrix $D = \{ d_{11}, d_{12}, ..., d_{nn} \}$ and Incidence Matrix $C = \{ c_{11}, c_{12}, ..., c_{nn} \}$.

Correlation $r$ between $D$ and $C$ is given by

$$r = \frac{\sum_{i=1, j=1}^{n} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i=1, j=1}^{n} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i=1, j=1}^{n} (c_{ij} - \bar{c})^2}}$$
Are There Clusters in the Data?

Random Points

K-means

DBSCAN

Complete Link
Measuring Cluster Validity Via Correlation

- Correlation of incidence and distance matrices for the K-means clusterings of the following two data sets

![Diagram 1](Corr = -0.9235)

![Diagram 2](Corr = -0.5810)
Using Similarity Matrix for Cluster Validation

- Order the similarity matrix with respect to cluster labels and inspect visually.
Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp
Reliability of Clusters

- Need a framework to interpret any measure
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?

- Statistics provide a framework for cluster validity
  - The more “atypical” a clustering result is, the more likely it represents valid structure in the data
Statistical Framework for SSE

• Example
  – Compare SSE of 0.005 against three clusters in random data
  – SSE Histogram of 500 sets of random data points of size 100—lowest SSE is 0.0173

SSE = 0.005
Determine the Number of Clusters Using SSE

- SSE curve

Clustering of Input Data

SSE wrt K
Take-away Message

• What’s clustering?
• Why clustering is important?
• How to preprocess data and compute dissimilarity/similarity from data?
• What’s a good clustering solution?
• How to evaluate the clustering results?