Data preprocessing

Data set types

- Record
  - Tables
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Tabular Data

- A collection of records
  - Each record is characterized by a fixed set of attributes

<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>

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

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>Term</th>
<th>coach</th>
<th>play</th>
<th>score</th>
<th>win</th>
<th>lose</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
### Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

---

### Ordered Data

- Sequences of transactions

Items/Events

\[
(\text{A, B}) \rightarrow (\text{D}) \rightarrow (\text{C, E}) \\
(\text{B, D}) \rightarrow (\text{C}) \rightarrow (\text{E}) \\
(\text{C, D}) \rightarrow (\text{B}) \rightarrow (\text{A, E})
\]

An element of the sequence

---

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Attribute types

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

Examples of data quality problems:
- Noise and outliers
- missing values
- duplicate data

Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)
Important Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality

- Sparsity
  - Only presence counts

- Resolution
  - Patterns depend on the scale

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation
**Aggregation**

- Combining two or more attributes (or objects) into a single attribute (or object)

**Purpose**
- Data reduction
  - Reduce the number of attributes or objects
- Change of scale
  - Cities aggregated into regions, states, countries, etc
- More "stable" data
  - Aggregated data tends to have less variability

---

**Data reduction**

- It generates a reduced representation of the dataset. This representation is smaller in volume, but it can provide similar analytical results
  - sampling
    - It reduces the cardinality of the set
  - feature selection
    - It reduces the number of attributes
  - discretization
    - It reduces the cardinality of the attribute domain

---

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Sampling ...

- The key principle for effective sampling is the following:
  - using a sample will work almost as well as using the entire data sets, if the sample is representative
  - A sample is representative if it has approximately the same property (of interest) as the original set of data

Types of Sampling

- Simple Random Sampling
  - There is an equal probability of selecting any particular item

- Sampling without replacement
  - As each item is selected, it is removed from the population

- Sampling with replacement
  - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once

- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition
**Dimensionality Reduction**

- **Purpose:**
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- **Techniques**
  - Principle Component Analysis
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

---

**Discretization**

- **It** splits the domain of a continuous attribute in a set of intervals
  - It reduces the cardinality of the attribute domain

- **Techniques**
  - N intervals with the same width $W = (v_{\text{max}} - v_{\text{min}})/N$
    - Easy to implement
    - It can be badly affected by outliers and sparse data
    - Incremental approach
  - N intervals with (approximately) the same cardinality
    - It better fits sparse data and outliers
    - Non incremental approach
  - clustering
    - It well fits sparse data and outliers

---

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Discretization

Normalization

- It is a type of data transformation
  - The values of an attribute are scaled so as to fall within a small specified range, typically (-1,+1) or (0,+1)
- Techniques
  - min-max normalization
    \[ v' = \frac{v - \text{min}}{\text{max} - \text{min}} \times (\text{new}_{\text{max}} - \text{new}_{\text{min}}) + \text{new}_{\text{min}} \]
  - z-score normalization
    \[ v' = \frac{v - \text{mean}}{\text{stand}_{\text{dev}}} \]
  - decimal scaling
    \[ v' = \frac{v}{10^j} \]
    \( j \) is the smallest integer such that \( \max(|v'|) < 1 \)

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
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

- **Proximity** refers to a similarity or dissimilarity

---

**Euclidean Distance**

- **Euclidean Distance**

\[
\text{dist} = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}
\]

Where \(n\) is the number of dimensions (attributes) and \(p_k\) and \(q_k\) are, respectively, the \(k^{th}\) attributes (components) or data objects \(p\) and \(q\).

- Standardization is necessary, if scales differ.
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**

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

---

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 $q$.

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Similarity Between Binary Vectors

- Common situation is that objects, \( p \) and \( q \), have only binary attributes
- Compute similarities using the following quantities
  - \( M_{01} \): the number of attributes where \( p \) was 0 and \( q \) was 1
  - \( M_{10} \): the number of attributes where \( p \) was 1 and \( q \) was 0
  - \( M_{00} \): the number of attributes where \( p \) was 0 and \( q \) was 0
  - \( M_{11} \): the number of attributes where \( p \) was 1 and \( q \) was 1

- Simple Matching and Jaccard Coefficients
  - \( \text{SMC} = \frac{\text{number of matches}}{\text{number of attributes}} \)
    \[ = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}} \]
  - \( J \): number of 11 matches / number of not-both-zero attributes values
    \[ = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \]

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

SMC versus Jaccard: Example

\[ \begin{align*}
  p &= 10000000000 \\
  q &= 0000001001 \\
\end{align*} \]

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

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

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Cosine Similarity

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

- Example:
  \[
  d_1 = 3 2 0 5 0 0 0 2 0 0
  d_2 = 1 0 0 0 0 0 0 1 0 2
  \]
  \( d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 1*0 + 0*0 + 0*2 = 5 \)
  \( ||d_1|| = (3*3 + 2*2 + 0*0 + 5*0 + 0*0 + 0*0 + 1*1 + 0*0 + 0*0) = 6.481 \)
  \( ||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1) = 2.245 \)
  \[
  \cos( d_1, d_2 ) = \frac{5}{6.481 \times 2.245} = 0.7150
  \]

Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the \( k^{th} \) attribute, compute a similarity, \( s_k \), in the range \([0, 1]\).
2. Define an indicator variable, \( \delta_k \), for the \( k^{th} \) attribute as follows:
   \[
   \delta_k = \begin{cases} 
   0 & \text{if the \( k^{th} \) attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the \( k^{th} \) attribute} \\
   1 & \text{otherwise}
   \end{cases}
   \]
3. Compute the overall similarity between the two objects using the following formula:
   \[
   \text{similarity}(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}
   \]

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Combining Weighted Similarities

- May not want to treat all attributes the same.
  - Use weights $w_k$ which are between 0 and 1 and sum to 1.

\[
similarity(p, q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}
\]

\[
distance(p, q) = \left( \sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}
\]

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006