Data preprocessing

Elena Baralis
Politecnico di Torino
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></th>
<th>team</th>
<th>coach</th>
<th>play</th>
<th>ball</th>
<th>score</th>
<th>game</th>
<th>n</th>
<th>wi</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>
<td></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>
<td></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>
<td></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>

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

- Examples: Generic graph and HTML Links

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

- Benzene Molecule: $C_6H_6$

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

- Sequences of transactions

Items/Events

- (A B) (D) (C E)
- (B D) (C) (E)
- (C D) (B) (A E)

An element of the sequence

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

- Genomic sequence data

```plaintext
GGTTCCGCCTTCAGCCCCCGCGCC
CGCAGGGCCCGCCCCGCCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTACAGCACGCGAAGCGC
TGGGCTGCTGCTCGGACCAGGG
```
Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean

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

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness:  =  ≠
  - Order:  <  >
  - Addition:  +  -
  - Multiplication:  *  /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties

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

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Noise refers to modification of original values

- Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen

---

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

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

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

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Important Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality

- Sparsity
  - Only presence counts

- Resolution
  - Patterns depend on the scale

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

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

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

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

Variation of Precipitation in Australia

Standard Deviation of Average Monthly Precipitation

Standard Deviation of Average Yearly Precipitation

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

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
The key principle for effective sampling is the following:

- using a sample will work almost as well as using the entire data set, if the sample is representative

- A sample is representative if it has approximately the same property (of interest) as the original set of data

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

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

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies.

- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful.

Randomly generate 500 points

Compute difference between max and min distance between any pair of points

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

Purpose
- Avoid curse of dimensionality
- 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
- Principal Component Analysis
- Singular Value Decomposition
- Others: supervised and non-linear techniques

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

- Goal is to find a projection that captures the largest amount of variation in data

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

- Another way to reduce dimensionality of data

- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid

- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA

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

- Techniques
  - Brute-force approach
    - Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches
    - Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches
    - Features are selected before data mining algorithm is run
  - Wrapper approaches
    - Use the data mining algorithm as a black box to find best subset of attributes

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

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes.

- Three general methodologies
  - Feature Extraction
    - domain-specific
  - Mapping Data to New Space
  - Feature Construction
    - combining features

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
Mapping Data to a New Space

- Fourier transform
- Wavelet transform

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
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 fits well sparse data and outliers

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

Data Equal interval width

Data

Equal interval width

Equal frequency

K-means

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

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: $x^k$, log$(x)$, $e^x$, $|x|$  
  - Standardization and Normalization

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
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
    \[ v' = \frac{v - \text{min}}{\text{max} - \text{min}} (\text{new}_{\text{max}} - \text{new}_{\text{min}}) + \text{new}_{\text{min}} \]
  - 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 \( \max(\{|v'|\}) < 1 \)
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

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
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>Nominal</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>Ordinal</td>
<td>$d = \frac{</td>
<td>p-q</td>
</tr>
<tr>
<td>Interval or Ratio</td>
<td>$d =</td>
<td>p - q</td>
</tr>
</tbody>
</table>

Table 5.1. Similarity and dissimilarity for simple attributes

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

- Euclidean Distance

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

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 \( q \).

- Normalization is necessary, if scales differ.

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

Distance Matrix

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

- Minkowski Distance is a generalization of Euclidean Distance

\[ dist = \left( \sum_{k=1}^{n} |p_k - q_k|^{r} \right)^{\frac{1}{r}} \]

Where \( r \) is a parameter, \( n \) is the number of dimensions (attributes) and \( p_k \) and \( q_k \) are, respectively, the kth attributes (components) of data objects \( p \) and \( q \).

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

- $r = 1$. City block (Manhattan, taxicab, $L_1$ norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors.

- $r = 2$. Euclidean distance

- $r \rightarrow \infty$. “supremum” ($L_{\max}$ norm, $L_{\infty}$ norm) distance.
  - This is the maximum difference between any component of the vectors.

- Do not confuse $r$ with $n$, i.e., all these distances are defined for any number of dimensions.
## Minkowski Distance

### Distance Matrix

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
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 \( q \).
Similarity Between Binary Vectors

- 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
  \( SMC = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{M_{11} + M_{00}}{M_{01} + M_{10} + M_{11} + M_{00}} \)
  \( J = \frac{\text{number of 11 matches}}{\text{number of not-both-zero attributes values}} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} \)

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006
SMC versus Jaccard: Example

\[ p = 1 0 0 0 0 0 0 0 0 0 \]
\[ q = 0 0 0 0 0 0 1 0 0 1 \]

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

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

\[ J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{0}{2 + 1 + 0} = 0 \]
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 norm of vector $d$.

- Example:

  $d_1 = 3 2 0 5 0 0 0 2 0 0$
  $d_2 = 1 0 0 0 0 0 0 1 0 2$

  \[
  d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
  \]

  \[
  \|d_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481
  \]

  \[
  \|d_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245
  \]

  \[
  \cos(d_1, d_2) = 0.3150
  \]
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:

   $\delta_k = \begin{cases} 
   0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\
   1 & \text{otherwise} 
   \end{cases}$

3. Compute the overall similarity between the two objects using the following formula:

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

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.

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

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

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