Data mining: data preprocessing

**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

---

**Tabular Data**

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Married</th>
<th>Male</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Married</td>
<td>Yes</td>
<td>50K</td>
</tr>
<tr>
<td>No</td>
<td>Single</td>
<td>Yes</td>
<td>50K</td>
</tr>
<tr>
<td>Yes</td>
<td>Married</td>
<td>Yes</td>
<td>70K</td>
</tr>
<tr>
<td>Yes</td>
<td>Single</td>
<td>Yes</td>
<td>70K</td>
</tr>
<tr>
<td>No</td>
<td>Married</td>
<td>Yes</td>
<td>70K</td>
</tr>
<tr>
<td>Yes</td>
<td>Married</td>
<td>No</td>
<td>70K</td>
</tr>
<tr>
<td>Yes</td>
<td>Married</td>
<td>Yes</td>
<td>70K</td>
</tr>
<tr>
<td>Yes</td>
<td>Married</td>
<td>Yes</td>
<td>70K</td>
</tr>
</tbody>
</table>

---

**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.

```
  Document 1: 3 0 5 0 2 6 0 2 0 2
  Document 2: 0 7 0 2 1 0 0 3 0 0
  Document 3: 0 1 0 0 1 2 2 0 3 0
```

---

**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>

---

**Graph Data**

- Examples: Generic graph and HTML Links
**Data mining: data preprocessing**

### Chemical Data
- Benzene Molecule: $C_6H_6$

![Chemical Data Diagram](Image)


### Ordered Data
- Sequences of transactions

<table>
<thead>
<tr>
<th>Items/Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A B)</td>
</tr>
<tr>
<td>(B D)</td>
</tr>
<tr>
<td>(C E)</td>
</tr>
<tr>
<td>(D)</td>
</tr>
<tr>
<td>(E)</td>
</tr>
<tr>
<td>(A E)</td>
</tr>
</tbody>
</table>

An element of the sequence


### Ordered Data
- Genomic sequence data

```
GTTCGGCCCTCCAGCCCCGGCC
CAGGCGGGCCGCCGCCCCTC
GAGAGGCCCAGCCTGGGCGGGG
GGGGGAGGCGGGCCGCCCGAG
CCACGTAGCCAGAGACCCAGTGC
CTTCGTGCCGCTTAGACCTGAA
GTTCCTAGGCCACGCACGGACAG
GCCAGTTAGACCCGAGGCCGCC
TGGGTCTCTGCTGGACCCAGGG
```


### Ordered Data
- Spatio-Temporal Data

![Spatio-Temporal Data Diagram](Image)

Jan

Average Monthly Temperature of land and ocean


### 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


### Properties of Attribute Values
- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $\neq$
  - Order: $<$, $>$
  - Addition: $+$, $-$
  - Multiplication: $\ast$, $/$
  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties


---

Elena Baralis
Politecnico di Torino
**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.

**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

**Noise**

- 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

**Outliers**

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

**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

---

Elena Baralis
Politecnico di Torino
Data mining: data preprocessing

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

Variation of Precipitation in Australia

Aggregation
- Standard Deviation of Average Monthly Precipitation
- Standard Deviation of Average Yearly Precipitation

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

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.

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.

Elena Baralis
Politecnico di Torino
Data mining: data preprocessing

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.

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.

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

Dimensionality Reduction: PCA
- Goal is to find a projection that captures the largest amount of variation in data.

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.

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
Data mining: data preprocessing

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

Mapping Data to a New Space

- Fourier transform
- Wavelet transform

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

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

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
  \[
  v' = \frac{v - v_{\text{min}}}{v_{\text{max}} - v_{\text{min}}} \cdot (\text{new max} - \text{new min}) + \text{new min}
  \]
- z-score normalization
  \[
  v' = \frac{v - \text{mean}}{\text{stand dev}}
  \]
- decimal scaling
  \[
  v' = \frac{v}{10^j} \quad j \text{ is the smallest integer such that } \max(|v'|) < 1
  \]
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

Similarity/Dissimilarity for Simple Attributes

\( \rho \) 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 = \frac{p - q}{p + q} ) if ( p = q ) ( d = 0 ) if ( p \neq q ) ( s = \frac{1}{2} ) if ( p = q ) ( s = 0 ) if ( p \neq q )</td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td>( d = \frac{p - q}{p + q} ) (values mapped to integers ( 0 ) to ( n-1 )), where ( n ) is the number of values ( s = 1 - \frac{d}{n-1} )</td>
<td></td>
</tr>
<tr>
<td>Interval or Ratio</td>
<td>( d = \frac{</td>
<td>p - q</td>
</tr>
</tbody>
</table>

Table 5.1. Similarity and dissimilarity for simple attributes

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 \( q_k \), respectively, are the \( k \)th attributes (components) or data objects \( p \) and \( q \).

- Standardization is necessary, if scales differ.

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

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

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

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” (\( L_{\infty} \) 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 all numbers of dimensions.

Elena Baralis
Politecnico di Torino
Data mining: data preprocessing

### Minkowski Distance

<table>
<thead>
<tr>
<th>p1</th>
<th>p2</th>
<th>p3</th>
<th>p4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

### Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.
  - \( d(p, q) \geq 0 \) for all \( p \) and \( q \) and \( d(p, q) = 0 \) only if \( p = q \). (Positive definiteness)
  - \( d(p, q) = d(q, p) \) for all \( p \) and \( q \). (Symmetry)
  - \( d(p, q) \leq d(p, r) + d(r, q) \) for all points \( p \), \( q \), and \( r \). (Triangle Inequality)
  - \( d(p, q) \) is the distance (dissimilarity) between points (data objects), \( p \) and \( q \).

- A distance that satisfies these properties is a metric.

### Common Properties of a Similarity

- Similarities, also have some well known properties.
  - \( s(p, q) = 1 \) (or maximum similarity) only if \( p = q \).
  - \( s(p, q) = s(q, p) \) for all \( p \) and \( q \). (Symmetry)

where \( s(p, q) \) is the similarity between points (data objects), \( p \) and \( q \).

### Cosine Similarity

\[
\text{Cosine Similarity} = \frac{p \cdot q}{\|p\| \cdot \|q\|} = \frac{\sum_{i=1}^{n} p_i q_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \cdot \sqrt{\sum_{i=1}^{n} q_i^2}}
\]

- If \( d_j \) and \( d'_j \) are two document vectors, then
  - \( \cos(d_j, d'_j) = \frac{(d_j \cdot d'_j)}{\|d_j\| \cdot \|d'_j\|} \),
    where \( \cdot \) indicates vector dot product and \( \| d \| \) is the norm of vector \( d \).

- Example:
  - \( d_j = 3 2 0 5 0 0 0 2 0 0 \)
  - \( d'_j = 1 0 0 0 0 0 0 1 0 2 \)

\[
\|d_j\| = (3^2 + 2^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2)^{1/2} = 6.02
\]

\[
\|d'_j\| = (1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{1/2} = 3.15
\]

\[
\cos(d_j, d'_j) = \frac{3205 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0}{6.02 \cdot 3.15} = 0.51
\]

- Elena Baralis
  - Politecnico di Torino
### 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, $e_k$, for the $k^{th}$ attribute as follows:
   \[ e_k = \begin{cases} 
   0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing value for the } k^{th} \text{ attribute} \\
   1 & \text{otherwise} 
   \end{cases} \]
3. Compute the overall similarity between the two objects using the following formula:
   \[ \text{similarity}(p,q) = \frac{\sum_{k=1}^{n} e_k s_k}{\sum_{k=1}^{n} e_k} \]

### 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.

\[
\text{similarity}(p,q) = \frac{\sum_{k=1}^{n} w_k e_k s_k}{\sum_{k=1}^{n} w_k e_k}
\]
\[
\text{distance}(p,q) = \left( \sum_{k=1}^{n} w_k (p_k - q_k)^2 \right)^{1/2}
\]