Chapter 2: Measurement and Data

Ming Li

Department of Computer Science and Technology
Nanjing University
Data mining: at a glance

The real world

Knowledge

Data mining

Representation

Data

?
What is data?

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity.

- Symbolic representation
- Measurement should reflect the given property.

Data entities in the symbolic space should reflect the true relationship of the objects in the real-world.
Are we mining qualified data?

Since we have no control over the data collection process, we have to be aware of the quality of the data we are mining

- Are the measurement procedures appropriate?
- If data transformation is applied, is it appropriate?
- Are the relationships between objects expressed?
- How to describe data quality?
Various data representations

- “Flatted” data
  - Data matrix or table. (attribute, value)

- Temporal and spatial data
  - (locations, value)

- Text
  - bag-of-words, n-grams, …

- Images and video
  - Color + texture + …, visual words, …

- Structured data
  - Graph, tree, …
An simple example

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Weight</th>
<th>LoveShopping?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>24</td>
<td>45</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>20</td>
<td>40</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>25</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>43</td>
<td>53</td>
<td>3</td>
</tr>
</tbody>
</table>

- Gender = 1, male; Gender = 0, female
- LoveShopping = 5, extremely love;
  LoveShopping = 3, love;
  LoveShopping = 1, don’t love it at all
Types of attribute scales (I)

- Nominal scale

The values of the attribute are only “Labels”, which is used to distinguish each other.

- Finite number of values
- No order information
- No algebraic operation can be conducted

E.g., \{1, 2, 3\} \sim \{\text{Red, Green, Blue}\} \sim \{\text{Milk, Bread, Coffee}\}
Types of attribute scales (II)

- Ordinal scale

The values of the attribute is to indicate certain ordering relationship resided in the attribute.

- Order is more important than value!

- No algebraic operation can be conducted except those related to sorting.

e.g., \{1, 2, 3\} ~ \{So-So, Good, Excellent\}

~ \{Irrelevant, Partially relevant, Relevant\}
Types of attribute scales (III)

• Numerical scale

The values of the attribute is to indicate the quantity of some predefined unit.

• There should be a basic unit, which can be transformed to another one.

• The value is how many copies of the basic unit

• Some algebraic operation can be conducted w.r.t the meaning of the attribute

  e.g.,  4 km = 4 * 1km

  4 km is twice as longer as 2 km
Types of attribute scales (IV)

- Numerical scale
  - Ratio scale
    One type of numerical scale with fixed origin
    e.g., Kelvin temperature: 0K = -273.15°C
  - Interval scale
    One type of numerical scale with arbitrary origin
    e.g., Celsius temperature, Fahrenheit temperature
Why talk about attribute types?

We have to make sure that:

- The underlying meanings (relationships) have been properly encoded.

  The relationships between the entities in the empirical system being studied should correspond to relationships between number in the numerical system

- Choose appropriate operation for different types of attribute, although the numerical (symbolic) representation appears to be similar.

  e.g., pain level : {1,2,6}, {3,4,5}
The data transformation

Raw data are sometimes transformed in order to:

- Adjust the skewed data to meet the requirement of some methods to be applied on these data
- To relieve the burden of the subsequent data mining algorithms
- Explore and visualize the data well.
Legitimate transformation of different types of attribute

- **Nominal scale:**
  
  One-to-one mappings (=)  
  e.g., $1 \rightarrow 4$

- **Ordinal scale:**
  
  Monotonic increasing (<)  
  e.g., $\{1, 2, 3\} \rightarrow \{2, 6, 10\}$

- **Ratio scale:**
  
  Multiplication (*)  
  e.g., $2 \rightarrow 20$

- **Interval scale:**
  
  Affine (*, +)  
  e.g., $2 \rightarrow 21$
Some basic transformations (I)

Normalization is to scale the (numerical) attribute values to some specified range

• Min-Max Normalization:

\[ v_{A'} = \frac{v_A - \min(A)}{\max(A) - \min(A)} (\max(A') - \min(A')) + \min(A') \]

• z-score normalization:

\[ v_{A'} = \frac{v_A - \mu_A}{\sigma_A} \]

• Decimal scaling normalization:

\[ v_{A'} = \frac{v_A}{10^j} \text{ where } j \text{ is the smallest integer such that } \max(|v'_{A}|) < 1 \]
Some basic transformations (II)

Discretization: dividing values of numerical attributes into intervals and use the interval to replace the original value

- **Entropy-based discretization:**

  Entropy of the sample $S$: $E(S) = - \sum_i p_i \log(p_i)$

  Entropy of the sample $S$ if split by boundary $b$: $I(S, b) = \frac{|S_1|}{|S|} E(S_1) + \frac{|S_2|}{|S|} E(S_2)$

  Criterion for split: $Gain(S, b) = E(S) - I(S, b) > \theta$

- **Discretization by natural partitioning**

  The 3-4-5 rule: For the most significant digit, if it covers $\{3, 6, 7, 9\}$ distinct values then divide it into 3 equi-width interval; if it covers $\{2, 4, 8\}$ distinct values then divide it into 4 equi-width interval; if it covers $\{1, 5, 10\}$ then divide it into 5 equi-width interval
Similarity: relationship between objects

• Similarity is important:
  – Represent the internal relationship between data objects.
  – It is essential to many data mining algorithms

• The similarities computed from data items should reflect the relationship between objects

• Why care about similarity?
  – Find appropriate similarity measure for certain data representation
  – Help to evaluate the quality of data representation
Distance measures

• Distance measure can be used to characterize the concept of “similarity”

• **Distance or Metric** should satisfy

  – **Non-negativity:** \( d(i, j) \geq 0 \) and \( d(i, j) = 0 \) iff \( i = j \)

  – **Symmetry:** \( d(i, j) = d(j, i) \) for all \( i, j \)

  – **Triangle inequality:** \( d(i, j) \leq d(i, k) + d(k, j) \) for all \( i, j \) and \( k \)
Widely-used distance (I)

Minkowski distance:

\[ d(x_1, x_2) = \left( \sum_{k=1}^{d} (x_1(k) - x_2(k))^p \right)^{\frac{1}{p}} \]

where \( x_1 \) and \( x_2 \) are 2 vectors in \( d \)-dimensional space

- Two special cases:
  - Euclidean distance (\( p = 2 \))
    \[ d(x_1, x_2) = \sqrt{\sum_{k=1}^{d} (x_1(k) - x_2(k))^2} \]
  - Manhattan distance (\( p = 1 \))
    \[ d(x_1, x_2) = \sum_{k=1}^{d} |x_1(k) - x_2(k)| \]

- Commensurability is assumed and normalization is usually required.
Widely-used distance (II)

Weighted Minkowski distance:

\[ d(x_1, x_2) = \left( \sum_{k=1}^{d} w_k (x_1(k) - x_2(k))^p \right)^{\frac{1}{p}} \]

Reflects the relative importance of each attribute

In both weighted and unweighted Minkowski distance, each attribute contribute independently to the measure of distance

Mahalanobis distance:

\[ d(x_1, x_2) = \left( (x_1 - x_2)^\top \Sigma^{-1} (x_1 - x_2) \right)^{\frac{1}{2}} \]

The covariance matrix

Mahalanobis distance standardizes data not only in the direction of each attributes but also based on the covariance between attributes
Widely-used distance (III)

- **Distance for binary data:**
  
  - **Hamming distance:** number of bits that are different
    
    e.g., 01010 01001  Dist = 2
  
  - **Matching coefficient for similarity measurement**
    
    \[
    Sim = \frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{0,0} + n_{1,0} + n_{0,1}}
    \]
  
  - **Jaccard coefficient:** *(matches on (0,0) is not important)*
    
    \[
    J = \frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}
    \]
  
  - **Dice coefficient:**
    
    \[
    D = \frac{2n_{1,1}}{2n_{1,1} + n_{1,0} + n_{0,1}}
    \]
Measuring similarity for other type

- **Nominal attributes**
  - Split it into $n$ different binary attribute
  - Applied VDM (value difference metric)

$$vdm_a(x, y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q$$

[Wilson & Martines, JAIR’97]

- **Complex structures**
  - For distribution: KL divergence, cross entropy, …
  - For trees, graphs: defining graph kernels, …
Learning distance metric from data

\[ d(x, y) = d_A(x, y) = \|x - y\|_A = \sqrt{(x - y)^T A (x - y)}. \]

\[
\begin{align*}
\min_A & \quad \sum_{(x_i, x_j) \in S} \|x_i - x_j\|^2_A \\
\text{s.t.} & \quad \sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A \geq 1, \\
& \quad A \succeq 0.
\end{align*}
\]

I don’t know how to measure distance, but I know which ones should be close to each other

We can learn the distance relationship from data

Reprinted from [Xing et al., NIPS02]
Data quality

• Why care about the data quality?
  – Data mining process has no control over the data collection process
  – We don’t want the discovered pattern to reflect the distorted properties of the data set rather than the ground-truth.

• How to characterize data quality?
  – Data quality for individual attribute
  – Data quality of the entire data set
Data quality of measure

- **Precise (reliable)**
  - A precise measurement procedure is one that has *small variability*
  - often measured by its *variance*
  - can be improved as the sample size increases

- **Accurate**
  - An accurate measurement procedure not only possesses *small variability*, but also yields results *close to* what we think of as the *true value*.
  - measured by both *bias* and *variance*

- **Valid**
  - A valid measurement procedure measures what it is supposed to measure.
Bias vs. Variance

The value

# times of sample

True value

bias

variance

# times of sample
The accurate measure

The value

# times of sample

True value
Data quality for entire data set

A high quality data set should lead to an accurate estimate of parameters

• Low variance
  
  Can be achieved by increasing sample size

• Low bias
  
  Closely related to the sampling process
  
  • Convenience sample: may distort the sample distribution
  
  • Population drift: collected data become useless quickly.
Let’s move to Chapter 3