Data Preprocessing
Need to Preprocess Data

- Data quality is a key issue with data mining

- To increase the accuracy of the mining, has to perform data preprocessing.
  - Otherwise, garbage in => garbage out

- 80% of mining efforts often spend their time on data quality
How to Preprocess Data?

- Data Cleaning
- Data Integration
- Data Normalization
- Data Reduction
Why Data Cleaning?

Real-world data are:

» Incomplete:
  – missing values, missing attributes, or containing only aggregate data

» Noisy:
  – containing errors or outliers

» Inconsistent:
  – containing discrepancies in codes or names

Solution: Data Cleaning
Why Data Integration?

- Data comes from different Sources with
  - Same concept but different attribute name:
    - (Example: ssn ; social_security ; student_ssn)
  - Same value expressed differently:
    - (Example: undergraduate ; UG…)
  - Repeated tuples in different source databases.

=> Causes inconsistencies and redundancies.

- Solution: Data Integration (schema re-consolidation)
Why Data Reduction?

- Huge amount of data
  - decreases the efficiency
  - Make analysis difficult

- Solution: Data Reduction (reducing huge dataset to smaller representation that can show the same analysis)
Why Data Normalization?

- The range of attributes (features) values differ, thus one feature might overpower the other one.

- Solution: Normalization (Scaling data values in a range such as [0..1], [-1..1]) prevents outweighing features with large range like ‘salary’ over features with smaller range like ‘age’.
Data Cleaning:
Handling Missing Values

- Use attribute mean.
- Use attribute mean for all samples belonging to the same class.
- Use most probable value based on existing data (via Decision Tree, Bayesian,...).
  ex.: What would probably be the salary of a person with age $x$ and education $y$ based on the other data we currently have?
- As these are all estimates, they can lead to invalid results!
Data Cleaning: Detect Noisy Data

- **Histogram** - data distribution analysis
- **Cluster Analysis** - by detecting data that are outside any cluster.
- **Regression** - by using regression function.
Data Cleaning: Smoothing Noisy Data

- **Binning** - by arranging the data into buckets.

- **Concept Hierarchy**
  - Example: presenting numeric values such as age as young, middle age, and old.

- **Ignoring outliers detected by**
  (Outliers are data that are outside of the range of or inconsistent with the remaining data)
    - Histogram
    - Clustering
    - Regression
Binning (Example)

- Step 1: Partition sorted values into equal size bins.
- Step 2: Smooth by bin means/medians/boundaries.
- \( \Rightarrow \) reduces distinct values and gets rid of outliers:
  - 4, 8, 15, 21, 21, 24, 25, 28, 34
    - Bin 1: 4, 8, 15
    - Bin 2: 21, 21, 24
    - Bin 3: 25, 28, 34
  - By bin mean:
    - Bin 1: 9, 9, 9; bin 2: 22, 22, 22; bin 3: 29, 29, 29
  - Smoothing by bin boundary
    - Bin 1: 4, 4, 15; bin 2: 21, 21, 24; bin 3: 25, 25, 34
Clustering

- Find clusters and look for elements outside of any cluster.
Regression

- Find “best fitting” curve to existing data points.
- Points not matching curve are outliers.

Example:
y = x is best fitting curve
For current data. The outliers are the three points outside of the curve.
Data Cleaning: Handling Inconsistent Data

- Using known Functional dependencies
  - (example: item# → item)

- Revisiting data integration, as some inconsistencies might exist because of different names of the same attribute.
Data Integration

- Consolidate different source into one repository, usually data warehouse (schema re-consolidation)
  - Using metadata
  - Correlation analysis (measure how strongly one attribute implies the other attribute).
Data Reduction

• To increase the efficiency, can reduce the huge data set to a smaller representative.

• Methods:
  » Data aggregation (data cubes)
    – example: number of items sold in year vs. in month.
  » Dimension/attribute reduction
  » Data Compression
  » Discretization
Discretization and Concept Hierarchy

- Discretization is to transform the numeric (Continues) data to Categorical values.
- Some data Mining Algorithms only accept categorical values.
- Example:
  - Continues data: 1,2,3,4,5,…,20
    - Discretized values: 1-5; 6-10; 11-15; 16-20
  - Continues data for feature Age: 1,…,99
    - Categorical values: 1-15 : assign this range to concept “child”
    - 16- 40 : assign this range to concept “Young”
    - and so on ………..
Data Normalization

- Scale the data value to a range using methods such as:
  - Min-Max
  - Z-Score
  - Decimal Scaling
Data Normalization: Min-Max

- Linear transformation of the original input range into a newly specified data range (typically 0-1).

\[ y' = \frac{y - \min}{\max - \min} (\max' - \min') + \min' \]

- Old min value is mapped to new min, \( \min' \).
- Old max is mapped to new max, \( \max' \).
- Let \( y \) be the original value, \( y' \) be the new value.
- \( \min, \max \) are the original min and max.
- \( \min', \max' \) are the new min and max.
Min-Max (Example)

Consider old data that ranged from 0-100, we now obtain an equation to migrate it to 5-10 range.

\[ y' = \frac{y - \text{min}}{\text{max} - \text{min}} (\text{max}' - \text{min}') + \text{min}' \]

- \( y' = (y/20) + 5 \)
- \( y = 0, \quad y' = 5 \)
- \( y = 10, \quad y' = 5.5 \)
- \( y = 90, \quad y' = 9.5 \)
Data Normalization: Z-Score

- useful when min and max are unknown or outliers dominate the value min-max.
- The goal is that most of the data will lie within the origin to a standard deviation.
- If majority of data falls within 50 and 100, but you have a few data points outside of that range, z-score will compress most of the data into a small range.

\[ y' = \frac{y - \text{mean}}{\text{std}} \]
<table>
<thead>
<tr>
<th>y</th>
<th>y'</th>
<th>Avg</th>
<th>std</th>
<th>y</th>
<th>y'</th>
<th>Avg</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.18</td>
<td>-0.84</td>
<td>0.68</td>
<td>0.59</td>
<td>20.00</td>
<td>-0.26</td>
<td>34.30</td>
<td>55.86</td>
</tr>
<tr>
<td>0.60</td>
<td>-0.14</td>
<td></td>
<td></td>
<td>40</td>
<td>0.11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.52</td>
<td>-0.27</td>
<td></td>
<td></td>
<td>65</td>
<td>0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>-0.72</td>
<td></td>
<td></td>
<td>70</td>
<td>0.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.80</td>
<td>0.20</td>
<td></td>
<td></td>
<td>32</td>
<td>-0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.55</td>
<td>-0.22</td>
<td></td>
<td></td>
<td>8</td>
<td>-0.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.92</td>
<td>0.40</td>
<td></td>
<td></td>
<td>5</td>
<td>-0.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.21</td>
<td>-0.79</td>
<td></td>
<td></td>
<td>15</td>
<td>-0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.64</td>
<td>-0.07</td>
<td></td>
<td></td>
<td>250</td>
<td>3.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.20</td>
<td>-0.80</td>
<td></td>
<td></td>
<td>32</td>
<td>-0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.63</td>
<td>-0.09</td>
<td></td>
<td></td>
<td>18</td>
<td>-0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.70</td>
<td>0.04</td>
<td></td>
<td></td>
<td>10</td>
<td>-0.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.67</td>
<td>-0.02</td>
<td></td>
<td></td>
<td>-14</td>
<td>-0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.58</td>
<td>-0.17</td>
<td></td>
<td></td>
<td>22</td>
<td>-0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.98</td>
<td>0.50</td>
<td></td>
<td></td>
<td>45</td>
<td>0.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.81</td>
<td>0.22</td>
<td></td>
<td></td>
<td>60</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.10</td>
<td>-0.97</td>
<td></td>
<td></td>
<td>-5</td>
<td>-0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.82</td>
<td>0.24</td>
<td></td>
<td></td>
<td>7</td>
<td>-0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>-0.30</td>
<td></td>
<td></td>
<td>2</td>
<td>-0.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.00</td>
<td>3.87</td>
<td></td>
<td></td>
<td>4</td>
<td>-0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>0.68</strong></td>
<td><strong>0.00</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Normalization: Decimal scaling

- divide the value by $10^n$ where $n$ is the number of digits of the maximum absolute value.

$$y' = \frac{y}{10^n}$$

» Example: $X=900$ is maximum value

$\Rightarrow n = 3$

$\Rightarrow 900$ scales to $0.009$. 
A main portion of Data Warehousing and Data Mining effort is to preprocess the data.

Data cleaning, integration, reduction, and normalization are used to preprocess the data.