ECT7110
Data Preprocessing

Prof. Wai Lam
Why Data Preprocessing?

• Data in the real world is dirty
  – incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  – noisy: containing errors or outliers
  – inconsistent: containing discrepancies in codes or names

• No quality data, no quality mining results!
  – Quality decisions must be based on quality data
  – Data warehouse needs consistent integration of quality data
Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

- **Data integration**
  - Integration of multiple databases, data cubes, or files

- **Data transformation**
  - Normalization and aggregation

- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results

- **Data discretization**
  - Part of data reduction but with particular importance, especially for numerical data
Forms of data preprocessing

Data Cleaning
- [water to clean dirty-looking data]
- [clean-looking data]
- [soap suds on data]

Data Integration

Data Transformation
-2, 32, 100, 59, 48
-0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction

ECT7110 Data Preprocessing 4
Data Cleaning

• Data cleaning tasks
  – Fill in missing values
  – Identify outliers and smooth out noisy data
  – Correct inconsistent data
Data Transformation

• Smoothing: remove noise from data
• Aggregation: summarization, data cube construction
• Normalization: scaled to fall within a small, specified range
  – min-max normalization
  – z-score normalization
  – normalization by decimal scaling
• Attribute/feature construction
  – New attributes constructed from the given ones
Data Transformation: Normalization

• min-max normalization

\[ v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A}(\text{new}_\text{max}_A - \text{new}_\text{min}_A) + \text{new}_\text{min}_A \]

• z-score normalization

\[ v' = \frac{v - \text{mean}_A}{\text{stand}_\text{dev}_A} \]
Normalization - Examples

• Suppose that the minimum and maximum values for attribute income are 12,000 and 98,000 respectively. How to map an income value of 73,600 to the range of [0.0,1.0]?

• Suppose that the mean and standard deviation for the attribute income are 54,000 and 15,000. How to map an income value of 73,600 using z-score normalization?
Data Reduction Strategies

• Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

• Data reduction
  – Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

• **Data reduction strategies**
  – Dimensionality reduction
Dimensionality Reduction

• Feature selection (i.e., attribute subset selection):
  – Select a minimum set of features useful for data mining
  – reduce # of patterns in the patterns, easier to understand
Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Related to quantization problems.
Histograms

Singleton buckets

Buckets denoting a Continuous range of values
Histograms

• How are buckets determined and the attribute values partitioned?
  - Equiwidth: The width of each bucket range is uniform
  - Equidepth: The buckets are created so that, roughly, the frequency of each bucket is constant
Histogram Examples

• Suppose that the values for the attribute *age*:

13, 15, 16, 16, 19, 20, 20, 21, 21, 22, 25, 25, 25, 25, 30, 30, 30, 30, 32, 33, 33, 37, 40, 40, 40, 42, 42

**Equiwidth Histogram:**

<table>
<thead>
<tr>
<th>Bucket range</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>13-22</td>
<td>10</td>
</tr>
<tr>
<td>23-32</td>
<td>9</td>
</tr>
<tr>
<td>33-42</td>
<td>8</td>
</tr>
</tbody>
</table>

**Equidepth Histogram:**

<table>
<thead>
<tr>
<th>Bucket range</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>13-21</td>
<td>9</td>
</tr>
<tr>
<td>22-30</td>
<td>9</td>
</tr>
<tr>
<td>32-42</td>
<td>9</td>
</tr>
</tbody>
</table>
Sampling

- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data

- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew

- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

- Sampling may not reduce database I/Os (page at a time).
Sampling

**SRSWOR**
(simple random sample without replacement)

**SRSWR**
Cluster sample (m 2)

Stratified sample (according to age)

ECT7110 Data Preprocessing