Data Mining MTAT.03.183
Descriptive analysis, preprocessing, visualisation...

Jaak Vilo
2016 Spring
Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process
Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - noisy: containing errors or outliers
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records
Why Is Data Dirty?

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems

- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission

- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)

- Duplicate records also need data cleaning
Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

- Garbage in, garbage out
High quality data requirements

- High-quality data needs to pass a set of quality criteria. Those include:

  - **Accuracy**: an aggregated value over the criteria of integrity, consistency, and density
  - **Integrity**: an aggregated value over the criteria of completeness and validity
  - **Completeness**: achieved by correcting data containing anomalies
  - **Validity**: approximated by the amount of data satisfying integrity constraints
  - **Consistency**: concerns contradictions and syntactical anomalies
  - **Uniformity**: directly related to irregularities and in compliance with the set 'unit of measure'
  - **Density**: the quotient of missing values in the data and the number of total values ought to be known

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2.75</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>NA</td>
<td>5</td>
<td>4.75</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>3.25</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2.25</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>4.75</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2.66666667</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>3.44444444</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2.25</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>4</td>
<td>5.25</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.25</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>4.5</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>na</td>
<td>na</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1.75</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3.25</td>
</tr>
<tr>
<td>mean</td>
<td>3.76470588</td>
<td>2.75</td>
<td>3.0625</td>
<td>3.31578947</td>
<td>3.23529412</td>
</tr>
</tbody>
</table>
Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility

- Broad categories:
  - Intrinsic, contextual, representational, and accessibility
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

Data Integration

Data Transformation

Data Reduction
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary
| 0.48 0.03 0.06 0.05 0.43 0.19 0.16 0.35 0.25 0.07 |
| 0.29 0.14 0.96 0.02 0.11 0.22 0.80 0.05 0.54 0.36 |
| 0.23 0.28 0.02 0.10 0.48 0.31 0.36 0.21 0.33 0.45 |
| 0.64 0.04 0.48 0.56 0.16 0.58 0.33 0.11 0.42 0.06 |
| 0.00 0.23 0.24 0.00 0.54 0.02 0.26 0.20 0.18 0.01 |
| 0.17 0.17 0.04 0.97 0.25 0.04 0.34 0.01 0.50 0.15 |
| 0.43 0.05 0.50 0.16 0.52 0.82 0.23 0.09 0.02 0.21 |
| 0.13 0.17 0.33 0.26 0.00 0.33 0.57 0.43 0.09 0.43 |
| 0.24 0.08 0.08 0.54 0.08 0.02 0.02 0.01 0.35 0.62 |
| 0.10 0.03 0.14 0.78 0.30 0.07 0.08 0.48 0.57 0.30 |
Characterise data

use Big_University_DB
mine characteristics as "Science_Students"
in relevance to
name, gender, major, birth_date, residence, phone#, gpa
from student
where status in graduate
Mining Data Descriptive Characteristics

- **Motivation**
  - To better understand the data: central tendency, variation and spread

- **Data dispersion characteristics**
  - median, max, min, quantiles, outliers, variance, etc.

- **Numerical dimensions** correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals

- **Dispersion analysis on computed measures**
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube
Measuring the Central Tendency

- **Mean (algebraic measure) (sample vs. population):**
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \mu = \frac{\sum x}{N} \]
  - Weighted arithmetic mean:
  \[ \bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \]
  - Trimmed mean: chopping extreme values

- **Median:** A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):
  \[ \text{median} = L_1 + \left( \frac{n/2 - \left( \sum f \right)l}{f_{\text{median}}} \right)c \]

- **Mode**
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: \[ \text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median}) \]
- **Histograms and Probability Density Functions**
  - **Probability Density Functions**
    - Total area under curve integrates to 1
  - **Frequency Histograms**
    - *Y-axis is counts*
    - Simple interpretation
    - Can't be directly related to probabilities or density functions
  - **Relative Frequency Histograms**
    - Divide counts by total number of observations
    - *Y-axis is relative frequency*
    - Can be interpreted as probabilities for each range
    - Can't be directly related to density function
      - Bar heights sum to 1 but won't integrate to 1 unless bar width = 1
  - **Density Histograms**
    - Divide counts by (total number of observations * bar width)
    - *Y-axis is density values*
    - Bar height * bar width gives probability for each range
    - Can be directly related to density function
      - Bar areas sum to 1

[http://www.geog.ucsb.edu/~joel/g210_w07/lecture_notes/lect04/oh07_04_1.html](http://www.geog.ucsb.edu/~joel/g210_w07/lecture_notes/lect04/oh07_04_1.html)
histograms

• equal sub-intervals, known as `bins`

• break points

• bin width
The data are (the log of) wing spans of aircraft built in from 1956 - 1984.

Histogram with breaks at n.0 and n.5
binwidth=0.5
Histogram with breaks at n.25 and n.75
binwidth=0.5
2-dimensional data: dot-plot
Histogram vs kernel density

• properties of histograms with these two examples:
  – they are not smooth
  – depend on end points of bins
  – depend on width of bins

• We can alleviate the first two problems by using kernel density estimators.

• To remove the dependence on the end points of the bins, we centre each of the blocks at each data point rather than fixing the end points of the blocks.
'Histogram' with blocks centred over data points

block of width 1 and height 1/12 (the dotted boxes) as they are 12 data points, and then add them up
• Blocks - it is still discontinuous as we have used a discontinuous kernel as our building block
• If we use a smooth kernel for our building block, then we will have a smooth density estimate.
• It's important to choose the most appropriate bandwidth as a value that is too small or too large is not useful.

• If we use a normal (Gaussian) kernel with bandwidth or standard deviation of 0.1 (which has area 1/12 under the each curve) then the kernel density estimate is said to undersmoothed as the bandwidth is too small in the figure below.

• It appears that there are 4 modes in this density - some of these are surely artifices of the data.
• Choose optimal bandwidth
  – Methods to estimate it

• AMISE = Asymptotic Mean Integrated Squared Error
• optimal bandwidth = arg min AMISE
• The optimal value of the bandwidth for our dataset is about 0.25.

• From the optimally smoothed kernel density estimate, there are two modes. As these are the log of aircraft wing span, it means that there were a group of smaller, lighter planes built, and these are clustered around 2.5 (which is about 12 m).

• Whereas the larger planes, maybe using jet engines as these used on a commercial scale from about the 1960s, are grouped around 3.5 (about 33 m).
• The properties of kernel density estimators are, as compared to histograms:
  – smooth
  – no end points
  – depend on bandwidth
Kernel Density estimation - links

- Tutorial and Java applet for testing:
- Simple interactive web demo at U. Tartu:
  - http://kde.tume-maailm.pri.ee/

Jaak Vilo and other authors
Definition

If \( x_1, x_2, ..., x_n \sim f \) is an independent and identically-distributed sample of a random variable, then the kernel density approximation of its probability density function is

\[
\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right)
\]

where \( K \) is some kernel and \( h \) is a smoothing parameter called the bandwidth. Quite often \( K \) is taken to be a standard Gaussian function with mean zero and variance 1. Thus the variance is controlled indirectly through the parameter \( h \):

\[
K \left( \frac{x - x_i}{h} \right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}.
\]
R – example (due K. Tretjakov)

d = c(1,2,2,2,2,1,2,2,2,2,3,2,3,4,5,4,3,2,3,4,4,5,6,7);

kernelsmooth <- function(data, sigma, x) {
  result = 0;
  for (d in data) {
    result = result + exp(-(x-d)^2/2/sigma^2);
  }
  result/sqrt(2*pi)/sigma;
}

x = seq(min(d), max(d), by=0.1);
y = sapply(x, function(x) { kernelsmooth(d, 1, x) });
hist(d);
lines(x,y);
1-Dimensional Distributions

Click to add points

Bandwidth (width of kernel)

- Set BW: 0.5
- Automatic BW

Bandwidth selector: Local 4 NN BWP1.0
Bandwidth factor: 1.00
Number of neighbours: 4

Kernel type: Uniform

Estimated distribution

Entropy of distribution = 6.11

point (6.33; 1.5) 20 points
1-Dimensional Distributions

Click to add points

Bandwidth (width of kernel)
- Set BW: 0.5
- Automatic BW

Bandwidth selector: Local 4-NN BW 1.0
Bandwidth factor: 1.00
Number of neighbours: 4

Kernel type: Gaussian

Estimated distribution

Entropy of distribution = 6.31

point (5.33; 1.6) 20 points
http://kde.tume-maailm.pri.ee/
http://kde.tume-maailm.pri.ee/
Kernel Density Estimation

Dataset
Click to add / remove points

Kernel
Generate CSV Settings Stats

Kernel Density Estimation by Raimond Tunnel, Lauri Hämari is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Jaak Vilo and other authors  UT: Data Mining 2009
Aggregation, analysis and visualization of geodata

- http://sightsmap.com
- Several large crowd-sourced datasets:
  - The whole Panoramio photobank used by Google maps
  - The whole Wikipedia, geotags and wikipedia article logs
  - Foursquare
  - ... More
- Aggregate data, calculate popularites of places, calculate type tags
- Translate and categorize titles and descriptions to get types
- Improve aggregation algorithms by learning
- Visualize heatmaps, type tags, aggregated sources

By: Tanel Tammet
By: Tanel Tammet
By: Tanel Tammet
non-parametric density estimation

P(x₁, x₂ | ωₙ)
• [http://176.32.89.45/~hideaki/res/kernel.html](http://176.32.89.45/~hideaki/res/kernel.html)
MDL Histogram Density Estimation

Petri Kontkanen, Petri Myllymäki
Complex Systems Computation Group (CoSCo)
Helsinki Institute for Information Technology (HIIT)
University of Helsinki and Helsinki University of Technology
P.O.Box 68 (Department of Computer Science)
FIN-00014 University of Helsinki, Finland
{Firstname}.{Lastname}@hiit.fi

Abstract

We regard histogram density estimation as a model selection problem. Our approach is based on the information-theoretic minimum description length (MDL) principle, which can be applied for tasks such as data clustering, density estimation, image denoising and model selection in general. MDL-based model selection is formalized via the normalized maximum likelihood (NML) distribution, which has several desirable optimality properties. We show how this frame-
only on finding the optimal bin count. These regular histograms are, however, often problematic. It has been argued (Rissanen, Speed, & Yu, 1992) that regular histograms are only good for describing roughly uniform data. If the data distribution is strongly non-
uniform, the bin count must necessarily be high if one wants to capture the details of the high density portion of the data. This in turn means that an unnecessary large amount of bins is wasted in the low density region.

To avoid the problems of regular histograms one must allow the bins to be of variable width. For these irregular histograms, it is necessary to find the optimal set
Figure 2: The Gaussian finite mixture densities $gm6$ and $gm8$ and the NML-optimal histograms with sample size 10000.
The diagram shows a comparison of different density and bandwidth models. The grey bars represent the histogram, while the blue line represents the fixed (optimized) model, and the red line represents the locally adaptive model. The x-axis and y-axis labels are not explicitly mentioned in the image.
• R tutorial
  – http://cran.r-project.org/doc/manuals/R-intro.html
  – http://www.google.com/search?q=R+tutorial
More links on R and kernel density

- [http://sekhon.berkeley.edu/stats/html/density.html](http://sekhon.berkeley.edu/stats/html/density.html)
- [http://www.google.com/search?q=kernel+density+estimation+R](http://www.google.com/search?q=kernel+density+estimation+R)
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Id</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>6</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>19</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>20</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>21</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>22</td>
<td>6</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>23</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>24</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>25</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>26</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>27</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>28</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>29</td>
<td>6</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>30</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>31</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>32</td>
<td>5</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>33</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Excel Sheet](Example_Data_Scatterplots.xlsx)
Exercise 2

In the plot below all four continuous variables are plotted against each other and the fifth variable - species - indicates the color (pink - setosa, green - versicolor, blue - virginica). Setosa species are recognizable as their petal width and length are very small. Versicolor and virginica species can be distinguished as petal width and length is a bit smaller and also sepal width and length is rather smaller for versicolors than virginicas.
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: \( Q_1 \) (25\(^{th} \) percentile), \( Q_3 \) (75\(^{th} \) percentile)
  - Inter-quartile range: \( \text{IQR} = Q_3 - Q_1 \)

- **Five number summary**: min, \( Q_1 \), M, \( Q_3 \), max
- **Boxplot**: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
- **Outlier**: usually, a value higher/lower than 1.5 \( \times \) IQR

- Variance and standard deviation (**sample**: \( s \), **population**: \( \sigma \))
  - **Variance**: (algebraic, scalable computation)
    \[
    s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \sum_{i=1}^{n} x_i \right]^2
    \]
    \[
    \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2
    \]
  - **Standard deviation** \( s \) (or \( \sigma \)) is the square root of variance \( s^2 \) (or \( \sigma^2 \))
Boxplot Analysis

- **Five-number summary** of a distribution:
  
  Minimum, Q1, M, Q3, Maximum

- **Boxplot**
  
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum
Box Plots

• Tukey77: John W. Tukey, "Exploratory Data Analysis". Addison-Wesley, Reading, MA. 1977.
  • [http://informationandvisualization.de/blog/box-plot](http://informationandvisualization.de/blog/box-plot)
1.5 x IQR – Inter Quartile range
Visualization of Data Dispersion: Boxplot Analysis
A Box Plot can show the difference in variance between replicates
Variance regularization can remove the bias
Violin plot
Violin plot – R
vertical kernel density

http://gallery.r-enthusiasts.com/graph/Violin_plot,43
Combining plots...

VANUS per AMET

- ETTEV: 44.8
- JUHT: 47.5
- MUU: 49.6
- PENSION: 38.1
- OPILANE: 73.3
- SPETS: 47.7
- TEENINDT: 44.6
- TÖÖLINE: 45.8
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements ($\mu$: mean, $\sigma$: standard deviation)
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it
Example of plots-containing article:

http://www.kgs.ku.edu/Magellan/WaterLevels/CD/Reports/OFR04_57/rep00.htm
Quantile-Quantile (q-q) Plots

- [http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html](http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html)
Cumulative Distribution Function (CDF)
A Q–Q plot comparing the distributions of standardized daily maximum temperatures at 25 stations in the US state of Ohio in March and in July. The curved pattern suggests that the central quantiles are more closely spaced in July than in March, and that the March distribution is skewed to the right compared to the July distribution. The data cover the period 1893–2001.
Parametric modeling usually involves making assumptions about the shape of data, or the shape of residuals from a regression fit. Verifying such assumptions can take many forms, but an exploration of the shape using histograms and q-q plots is very effective. The q-q plot does not have any design parameters such as the number of bins for a histogram.
Randomized expression data

Yeast 2-hybrid studies

Known (literature) PPI

MPK1 YLR350w
SNF4 YCL046W
SNF7 YGR122W.
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane
Positively and Negatively Correlated Data
Weight

Diameter

Abalone data set – P. Danelson hw.

94
Bias vs Variance

Fig. 1 Graphical illustration of bias and variance.

Scott Fortmann-Roe
http://scott.fortmann-roe.com/docs/BiasVariance.html
Not Correlated Data
Pearson Correlation

\[ \rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \]

\[ \rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2} \sqrt{E(Y^2) - E(Y)^2}}. \]

\[ r = r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}. \]
Several sets of (x, y) points, with the Pearson correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.
### Numerical summary?

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of $x$ in each case</td>
<td>11 exact</td>
</tr>
<tr>
<td>Variance of $x$ in each case</td>
<td>10 exact</td>
</tr>
<tr>
<td>Mean of $y$ in each case</td>
<td>7.50 (to 2 d.p.)</td>
</tr>
<tr>
<td>Variance of $y$ in each case</td>
<td>3.75 (to 2 d.p.)</td>
</tr>
<tr>
<td>Correlation between $x$ and $y$ in each case</td>
<td>0.816 (to 3 d.p.)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 d.p. and 3 d.p. resp.)</td>
</tr>
</tbody>
</table>
Anscombe’s quartet

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of $x$ in each case</td>
<td>11 exact</td>
</tr>
<tr>
<td>Variance of $x$ in each case</td>
<td>10 exact</td>
</tr>
<tr>
<td>Mean of $y$ in each case</td>
<td>7.50 (to 2 d.p.)</td>
</tr>
<tr>
<td>Variance of $y$ in each case</td>
<td>3.75 (to 2 d.p.)</td>
</tr>
<tr>
<td>Correlation between $x$ and $y$ in each case</td>
<td>0.816 (to 3 d.p.)</td>
</tr>
<tr>
<td>Linear regression line in each case</td>
<td>$y = 3.00 + 0.500x$ (to 2 d.p. and 3 d.p. resp.)</td>
</tr>
</tbody>
</table>
### Anscombe's Quartet

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x</td>
<td>y</td>
<td>x</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence.
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression.
Graphic Displays of Basic Statistical Descriptions

- Histogram: (shown before)
- Boxplot: (covered before)
- Quantile plot: each value $x_i$ is paired with $f_i$, indicating that approximately $100 f_i \%$ of data are $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary
Data Cleaning

- Importance
  - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
  - “Data cleaning is the number one problem in data warehousing”—DCI survey

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration
Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.
How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.

- Fill in the missing value manually: tedious + infeasible?

- Fill in it automatically with
  - a global constant: e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter

  - **the most probable value: inference-based such as Bayesian formula or decision tree**
K-NN impute

- K nearest neighbours imputation
- Find K neighbours on available data points
- Estimate the missing value

(Hastie, Tibshirani, Troyanskaya, ... Stanford 1999-2001)
<table>
<thead>
<tr>
<th>id</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2.75</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>NA</td>
<td>5</td>
<td>3.6666666</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>na</td>
<td>2</td>
<td>3.3333333</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>3.25</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2.25</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>4.75</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2.6666667</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>4</td>
<td>na</td>
<td>6</td>
<td>4.3333333</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1.25</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>4</td>
<td>1</td>
<td>12</td>
<td>5.25</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>7</td>
<td>na</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>15</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>4.5</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1.75</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>3.25</td>
</tr>
</tbody>
</table>

**Series Title**

- **Series1**
- **Series2**
- **Series3**
- **Series4**
- **Series5**
- **Series6**
- **Series7**
- **Series8**
- **Series9**
- **Series10**
- **Series11**
- **Series12**
- **Series13**
- **Series14**
- **Series15**
- **Series16**
- **Series17**
- **Series18**
- **Series19**
Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data
How to Handle Noisy Data?

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

- **Regression**
  - smooth by fitting the data into regression functions

- **Clustering**
  - detect and remove outliers

- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)
Simple Discretization Methods: Binning

- **Equal-width** (distance) partitioning
  - Divides the range into $N$ intervals of equal size: *uniform grid*
  - if $A$ and $B$ are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well

- **Equal-depth** (frequency) partitioning
  - Divides the range into $N$ intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky
Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34
Regression

\[ y = x + 1 \]
Cluster Analysis
Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter’s Wheels)
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary
Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id = B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units
Data preprocessing  [edit]

Record linkage is highly sensitive to the quality of the data being linked, so all data sets under consideration (particularly their key identifier fields) should ideally undergo a data quality assessment prior to record linkage. Many key identifiers for the same entity can be presented quite differently between (and even within) data sets, which can greatly complicate record linkage unless understood ahead of time. For example, key identifiers for a man named William J. Smith might appear in three different data sets as so:

<table>
<thead>
<tr>
<th>Data set</th>
<th>Name</th>
<th>Date of birth</th>
<th>City of residence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set 1</td>
<td>William J. Smith</td>
<td>1/2/73</td>
<td>Berkeley, California</td>
</tr>
<tr>
<td>Data set 2</td>
<td>Smith, W. J.</td>
<td>1973.1.2</td>
<td>Berkeley, CA</td>
</tr>
<tr>
<td>Data set 3</td>
<td>Bill Smith</td>
<td>Jan 2, 1973</td>
<td>Berkeley, Calif.</td>
</tr>
</tbody>
</table>

In this example, the different formatting styles lead to records that look different but in fact all refer to the same entity with the same logical identifier values. Most, if not all, record linkage strategies would result in more accurate linkage if these values were first normalized or standardized into a consistent format (e.g., all names are "Surname, Given name", all dates are "YYYY/MM/DD", and all cities are "Name, 2-letter state abbreviation"). Standardization can be accomplished through simple rule-based data transformations or more complex procedures such as lexicon-based tokenization and probabilistic hidden Markov models.[8]
Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality
NOTE: This project is in transition over to https://github.com/OpenRefine.

- New documentation wiki
- New issue tracking

Google Refine is a power tool for working with messy data, cleaning it up, transforming it from one format into another, extending it with web services, and linking it to databases like Freebase.
GitHub

OpenRefine

http://openrefine.org

openrefine.github.com

Github pages repository for OpenRefine account
Updated 24 days ago

OpenRefine

OpenRefine is a free, open source power tool for working with messy data and improving it
Updated on Feb 9

reconciliation_service_skeleton

Skeleton for Standalone Python Reconciliation Service for Google Refine
Updated on Apr 19, 2014

open-reconcile

Java ★ 1 ▼ 1
Welcome!

OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another; extending it with web services; and linking it to databases like Freebase.

Please note that since October 2nd, 2012, Google is not actively supporting this project, which has now been rebranded to OpenRefine. Project development, documentation and promotion is now fully supported by volunteers. Find out more about the history of OpenRefine and how you can help the community.

Using OpenRefine - The Book

Using OpenRefine, by Ruben Verborgh and Max De Wilde, offers a great introduction to OpenRefine. Organized by recipes with hands on examples, the book covers the following topics:

1. Import data in various formats
2. Explore datasets in a matter of seconds
3. Apply basic and advanced cell transformations
4. Deal with cells that contain multiple values
5. Create instantaneous links between datasets
6. Filter and partition your data easily with regular expressions
7. Use named-entity extraction on full-text fields to automatically identify topics
Normalisation

- Making data comparable...
Elements of microarray statistics

\[ M = \log_2 R - \log_2 G \]
\[ A = \frac{1}{2} \left( \log_2 R + \log_2 G \right) \]
Before and After Normalization

![Histograms](image1.png)

![Scatter Plots](image2.png)
Normalisation can be used to transform data.
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- **Data reduction**
- Discretization and concept hierarchy generation
- Summary
Data Reduction Strategies

- Why data reduction?
  - A database/data 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
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

- Data reduction strategies
  - Data cube aggregation:
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary
Discretization

- Three types of attributes:
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis
Discretization and Concept Hierarchy

- Discretization
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute

- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)
### E.g.

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-18 y</td>
<td>1580</td>
</tr>
<tr>
<td>19-65 y</td>
<td>8394</td>
</tr>
<tr>
<td>66-105</td>
<td>2700</td>
</tr>
</tbody>
</table>

### Vs

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-20</td>
<td>1780</td>
</tr>
<tr>
<td>21-40</td>
<td>4200</td>
</tr>
<tr>
<td>41-60</td>
<td>4767</td>
</tr>
<tr>
<td>61-105</td>
<td>2900</td>
</tr>
</tbody>
</table>
Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals
Example of 3-4-5 Rule

Step 1:
- Min = -$351
- Low (i.e., 5%-tile) = -$159
- profit
- High (i.e., 95%-tile) = $1,838
- Max = $4,700

Step 2:
- msd = 1,000
- Low = -$1,000
- High = $2,000

Step 3:
- (-$1,000 - 0)
- (0 - $1,000)
- ($1,000 - $2,000)

Step 4:
- (-$400 - 0)
- (0 - $1,000)
- ($1,000 - $2,000)
- ($2,000 - $5,000)

MsD = 1,000
Low = -$1,000
High = $2,000

Step 2:
- msd = 1,000
- Low = -$1,000
- High = $2,000

Step 3:
- (-$1,000 - 0)
- (0 - $1,000)
- ($1,000 - $2,000)

Step 4:
- (-$400 - 0)
- (0 - $1,000)
- ($1,000 - $2,000)
- ($2,000 - $5,000)
Example

-351,976.00 .. 4,700,896.50

MIN= -351,976.00
MAX=4,700,896.50
LOW = 5th percentile -159,876
HIGH = 95th percentile 1,838,761
msd = 1,000,000 (most significant digit)
LOW = -1,000,000 (round down) HIGH = 2,000,000 (round up)

3 value ranges
1. (-1,000,000 .. 0]
2. (0 .. 1,000,000]
3. (1,000,000 .. 2,000,000]

Adjust with real MIN and MAX
1. (-400,000 .. 0]
2. (0 .. 1,000,000]
3. (1,000,000 .. 2,000,000]
4. (2,000,000 .. 5,000,000]
Recursive …

1.1. (-400,000 .. -300,000 ]
1.2. (-300,000 .. -200,000 ]
1.3. (-200,000 .. -100,000 ]
1.4. (-100,000 .. 0 ]

2.1. (0 .. 200,000 ]
2.2. (200,000 .. 400,000 ]
2.3. (400,000 .. 600,000 ]
2.4. (600,000 .. 800,000 ]
2.5. (800,000 .. 1,000,000 ]

3.1. (1,000,000 .. 1,200,000]
3.2. (1,200,000 .. 1,400,000]
3.3. (1,400,000 .. 1,600,000]
3.4. (1,600,000 .. 1,800,000]
3.5. (1,800,000 .. 2,000,000]

4.1. (2,000,000 .. 3,000,000]
4.2. (3,000,000 .. 4,000,000]
4.3. (4,000,000 .. 5,000,000]
Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - \{Urbana, Champaign, Chicago\} < Illinois
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: \{street, city, state, country\}
Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set.
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy.
  - Exceptions, e.g., weekday, month, quarter, year

```
country                       15 distinct values
province_or_state            365 distinct values
  city                        3567 distinct values
    street                   674,339 distinct values
```
Chapter 2: Data Preprocessing

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary
Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining.
- Descriptive data summarization is need for quality data preprocessing.
- Data preparation includes:
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot of methods have been developed but data preprocessing still an active area of research.
References

- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD’02.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999