Data Mining MTAT.03.183

Descriptive analysis, preprocessing, visualisation...

Jaak Vilo
2011 Fall

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

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

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

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

Forms of Data Preprocessing

- Data cleaning
- Data integration
- Data transformation
- Data reduction
- Data discretization

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):
  \[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

- **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):
    \[ \text{median} = L + \left( \frac{n/2 - f_<}{f_{med}} \right) c \]

- **Mode**: Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:
    \[ \text{mode} = \text{mean} - 3 \times (\text{mean} - \text{median}) \]

Histograms and Probability Density Functions

- **Total area under curve integrates to 1**
- **Frequency histograms**
  - Simple to interpret
  - Can’t be directly related to probabilities or density functions

- **Relative frequency histograms**
  - Divide counts by total number of observations
  - Can’t be interpreted as probabilities for each range
  - Can’t be directly related to density functions

- **Density Histograms**
  - Divide counts by (total number of observations \times bar width)
  - Can be directly related to density function
  - Bar areas can’t

- **Bar heights sum to 1 but won’t integrate to 1 unless bar width = 1**
- **Density histograms**
  - Divide counts by (total number of observations \times bar width)
  - Can be directly related to density function
  - Bar areas sum to 1
The data are (the log of) wing spans of aircraft built in from 1956 - 1984.

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.

- 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 bandwith
  -- Need 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

• Gentle introduction
  – [Link to Gentle introduction](http://school.maths.uwa.edu.au/~duongt/seminars/intro2kde/)

• Tutorial
  – [Link to Tutorial](http://parallel.vub.ac.be/research/causalModels/tutorial/kde.html)

R – example (due K. Tretjakov)

```R
x = seq(min(d), max(d), by=0.1);
y = sapply(x, function(x) kernelsmooth(d, 1, x));
plot(x, y);
```

Jaak Vilo and other authors

UT: Data Mining 2009

More links on R and kernel density

- http://sekhon.berkeley.edu/stats/html/density.html

http://jmrl.csail.mit.edu/proceedings/papers/p2/kontkanen07a/kontkanen07a.pdf

MBL Histogram Density Estimation

Pekka Kontkanen, Pasi Myllymäki

University of Helsinki and Helsinki University of Technology
Finland

Abstract

We regard histogram density estimation as a useful selective problem. Our approach is based on the information-theoretic maximum likelihood principle, which can be applied for tasks such as data clustering, density estimation, image denoising and model selection in general. MBL-based model selection is facilitated via the maximized smoothed likelihood (MSL) deci- sion rule, which has shown plausible empirical properties, some of which show that it can only be found by the optimal bin count. These regu- larized histograms are, however, also problematic. It has been reported (Hannu, Spindler 
and Yu, 1992) that regular histograms are only good for smoothing roughly uniform data. If the data distribution is strong non- uniformity, the bin count must necessarily be high if one wants to capture the density of the high density portion of the data. This is even worse that on expiry it necessarily means that the density of the low density re- gion.

To avoid the problems of regular histograms one must allow the bins to be of variable width. For this reason we introduce a new method, it is necessary to find the optimal se-
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - Quartiles: Q₁ (25th percentile), Q₂ (50th percentile), Q₃ (75th percentile)
  - Inter-quartile range: IQR = Q₃ – Q₁

- Five number summary: min, Q₁, M, Q₃, 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 x IQR

- Variance and standard deviation (sample: s, population: σ²)
  - Variance: \( \sigma² = \frac{1}{n-1} \sum (x_i - \overline{x})² = \frac{1}{n-1} \sum x_i² - \left( \frac{1}{n-1} \sum x_i \right)² \)
  - Standard deviation \( s \) (or \( \sigma \)) is the square root of variance \( s² \) (or \( \sigma² \))

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
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

- [http://informationandvisualization.de/blog/box-plot](http://informationandvisualization.de/blog/box-plot)

Visualization of Data Dispersion: Boxplot Analysis

A Box Plot can show the difference in variance between replicates.
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.
- Allows the user to view whether there is a shift in going from one distribution to another.

http://en.wikipedia.org/wiki/Q-Q_plot

Kemmeren et al. (Mol. Cell, 2002)

- Randomized expression data
- Yeast 2-hybrid studies
- Known (literature) PPI

Protein/Gene pairs

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.
Not Correlated Data

Numerical summary?

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of $x$ in each case</td>
<td>11.0</td>
</tr>
<tr>
<td>Variance of $x$ in each case</td>
<td>10.0</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>Anscombe’s Quartet</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>$x$</td>
</tr>
<tr>
<td>10.0</td>
</tr>
<tr>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
</tr>
<tr>
<td>9.0</td>
</tr>
<tr>
<td>11.0</td>
</tr>
<tr>
<td>14.0</td>
</tr>
<tr>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>12.0</td>
</tr>
<tr>
<td>7.0</td>
</tr>
<tr>
<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 univariate 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)
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, 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 violations (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

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

Normalisation

- Making data comparable...

Elements of microarray statistics

\[ M = \log_2 R - \log_2 G \]
\[ A = \frac{1}{2} (\log_2 R + \log_2 G) \]
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

Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
- The aggregated data for an individual entity of interest
- E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
  - Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
  - Queries regarding aggregated information should be answered using data cube, when possible

Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - Further reduce the size of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

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

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)

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 equal-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

Example
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

References

- T. Dasu, T. Johnson, S. Mukkamala, 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