Data Preprocessing

Mirek Riedewald

Some slides based on presentation by
Jiawei Han and Micheline Kamber

Motivation

- Garbage-in, garbage-out
  - Cannot get good mining results from bad data
- Need to understand data properties to select the right technique and parameter values
- Data cleaning
- Data formatting to match technique
- Data manipulation to enable discovery of desired patterns

Data Records

- Data sets are made up of data records
- A data record represents an entity
- Examples:
  - Sales database: customers, store items, sales
  - Medical database: patients, treatments
  - University database: students, professors, courses
- Also called samples, examples, tuples, instances, data points, objects
- Data records are described by attributes
  - Database row = data record; column = attribute

Attributes

- Attribute (or dimension, feature, variable): a data field, representing a characteristic or feature of a data record
  - E.g., customerID, name, address
- Types:
  - Nominal (also called categorical)
    - No ordering or meaningful distance measure
  - Ordinal
    - Ordered domain, but no meaningful distance measure
  - Numeric
    - Ordered domain, meaningful distance measure
    - Continuous versus discrete

Attribute Type Examples

- Nominal: category, status, or "name of thing"
  - Hair_color = {black, brown, blond, red, auburn, grey, white}
  - marital status, occupation, ID numbers, zip codes
- Binary: nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known
  - Size = {small, medium, large}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- Interval
  - Measured on a scale of equal-sized units
  - Values have order
    - E.g., temperature in C or F, calendar dates
  - No true zero-point
- Ratio
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10m is twice as high as 5m).
    - E.g., temperature in Kelvin, length, counts, monetary quantities

Discrete vs. Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Nominal, binary, ordinal attributes are usually discrete
  - Integer numeric attributes
- Continuous Attribute
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Typically represented as floating-point variables

Data Preprocessing Overview

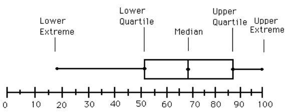
- Descriptive data summarization
- Data cleaning
- Data integration
- Data transformation
- Summary

Measuring the Central Tendency

- Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Trimmed mean: set weights of extreme values to zero
- Median
  - Middle value if odd number of values; average of the middle two values otherwise
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal distribution

Measuring Data Dispersion: Boxplot

- Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Various definitions for determining percentiles, e.g., for N records, the p-th percentile is the record at position (p/100)N+0.5 in increasing order
    - If not integer, round to nearest integer or compute weighted average
    - E.g., for N=30, p=25 (to get Q1): 25/100*30+0.5 = 8, i.e., Q1 is 8-th largest of the 30 values
    - E.g., for N=32, p=25: 25/100*32+0.5 = 8.5, i.e., Q1 is average of 8-th and 9-th largest values
- Boxplot: ends of the box are the quartiles, median is marked, whiskers extend to min/max
  - Often plots outliers individually
  - Outlier: usually, a value higher (or lower) than 1.5 x IQR from Q3 (or Q1)



Measuring Data Dispersion: Variance

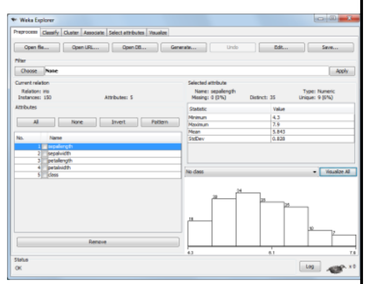
- Sample variance (aka second central moment):

$$m_2 = s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2$$

- Standard deviation = square root of variance
- Estimator of true population variance from a sample: $s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

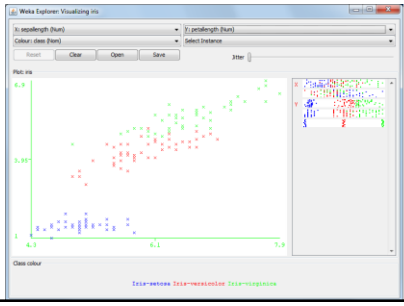
Histogram

- Graph display of tabulated frequencies, shown as bars
- Shows what proportion of cases fall into each category
- Area of the bar denotes the value, not the height
  - Crucial distinction when the categories are not of uniform width!

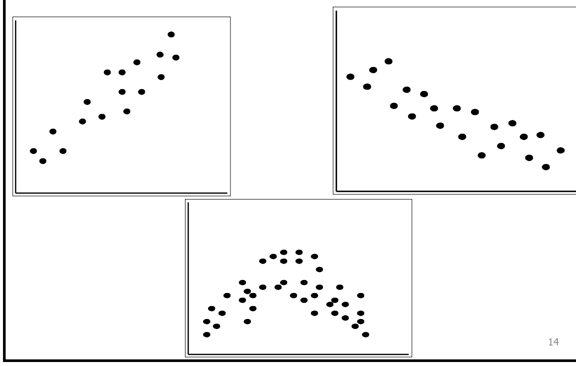


Scatter plot

- Visualizes relationship between two attributes, even a third (if categorical)
  - For each data record, plot selected attribute pair in the plane



Correlated Data



Not Correlated Data

Data Preprocessing Overview

- Descriptive data summarization
- Data cleaning
- Data integration
- Data transformation
- Summary

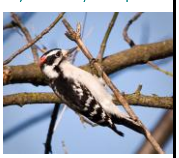
Why Data Cleaning?

- 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" and Birthday="03/07/1967"
    - E.g., was rating "1, 2, 3", now rating "A, B, C"
    - E.g., discrepancy between duplicate records

Example: Bird Observation Data

- Change of range boundaries over time, e.g., for temperature
- Different units, e.g., meters versus feet for elevation
- Addition or removal of attributes over the years
- Missing entries, especially for habitat and weather
  - People want to watch birds, not fill out long forms
- GIS data based on 30m cells or 1km cells
- Location accuracy
  - ZIP code versus GPS coordinates
  - Walk along transect but report only single location
- Inconsistent encoding of missing entries
  - 0, -9999, -3.4E+38—need context to decide
- Varying observer experience and capabilities
  - Confusion of species
  - Missed species that was present
- Confusion about reporting protocol
  - Report max versus sum seen
  - Report only interesting species, not all

Hairy vs. Downy Woodpecker



How to Handle Missing Data?

- Ignore the record
  - Usually done when class label is missing (for classification tasks)
- Fill in manually
  - Tedious and often not clear what value to fill in
- Fill in automatically with one of the following:
  - Global constant, e.g., “unknown”
    - “Unknown” could be mistaken as new concept by data mining algorithm
  - Attribute mean
  - Attribute mean for all records belonging to the same class
  - Most probable value: inference-based such as Bayesian formula or decision tree
    - Some methods, e.g., trees, can do this implicitly

How to Handle Noisy Data?

- Noise = random error or variance in a measured variable
- Typical approach: smoothing
  - Adjust values of a record by taking values of other “nearby” records into account
  - Dozens of approaches
    - Binning, average over neighborhood
    - Regression: replace original records with records drawn from regression function
    - Identify and remove outliers, possibly involving human inspection
- For this class: don’t do it unless you understand the nature of the noise
  - A good data mining technique should be able to deal with noise in the data

Data Preprocessing Overview

- Descriptive data summarization
- Data cleaning
- Data integration
- Data transformation
- Summary

Data Integration

- Combines data from multiple sources into a coherent store
- 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 might be different
  - Possible reasons: different representations, different scales, e.g., metric vs. US units
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
  - Can identify identical or similar attributes through **correlation analysis**

Covariance (Numerical Data)

- Covariance computed for data samples ($A_1, A_2,\ldots, A_n$) and ($B_1, B_2,\ldots, B_n$):

$$\mathrm{Cov}(A,B) = \frac{1}{n}\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B}) = \frac{1}{n}\sum_{i=1}^{n}A_iB_i - \bar{A}\cdot\bar{B}$$

- If A and B are independent, then Cov(A, B) = 0, but the converse is not true
  - Two random variables may have covariance of 0, but are not independent
- If Cov(A, B) > 0, then A and B tend to rise and fall together
  - The greater, the more so
- If covariance is negative, then A tends to rise as B falls and vice versa

Covariance Example

- Suppose two stocks A and B have the following values in one week:
  - A: (2, 3, 5, 4, 6)
  - B: (5, 8, 10, 11, 14)
  - AVG(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - AVG(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2·5+3·8+5·10+4·11+6·14)/5 − 4·9.6 = 4
- Cov(A,B) > 0, therefore A and B tend to rise and fall together

Correlation Analysis (Numerical Data)

- Pearson's product-moment correlation coefficient of random variables A and B: $\rho_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$
- Computed for two attributes A and B from data samples ($A_1, A_2, \ldots, A_n$) and ($B_1, B_2, \ldots, B_n$):

$$r_{A,B} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{A_i - \overline{A}}{s_A} \cdot \frac{B_i - \overline{B}}{s_B} \right)$$

Where $\overline{A}$ and $\overline{B}$ are the sample means, and s_A and s_B are the sample standard deviations of A and B (using the variance formula for s_n).

- Note: $-1 \le r_{A,B} \le 1$
  - $r_{A,B} > 0$: A and B positively correlated
    - The higher, the stronger the correlation
  - $r_{A,B} < 0$: negatively correlated

Correlation Analysis (Categorical Data)

- χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- The larger the χ^2 value, the more likely the variables are related
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-thefts in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- Numbers in parenthesis are expected counts calculated based on the data distribution in the two categories

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Data Preprocessing Overview

- Descriptive data summarization
- Data cleaning
- Data integration
- Data transformation
- Summary

Why Data Transformation?

- Make data more "mineable"
  - E.g., some patterns visible when using single time attribute (entire date-time combination), others only when making hour, day, month, year separate attributes
  - Some patterns only visible at right granularity of representation
- Some methods require normalized data
  - E.g., all attributes in range [0.0, 1.0]
- Reduce data size, both #attributes and #records

Normalization

- Min-max normalization to [new_min_A, new_max_A]:

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - E.g., normalize income range [$12,000, $98,000] to [0.0, 1.0]. Then $73,000 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- Z-score normalization (μ: mean, σ: standard deviation): $v' = \frac{v - \mu_A}{\sigma_A}$
  - E.g., for μ = 54,000 and σ = 16,000, $73,000 is mapped to $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$
- Normalization by decimal scaling: $v' = \frac{v}{10^j}$

where j is the smallest integer such that Max(|v'|) < 1

Data Reduction

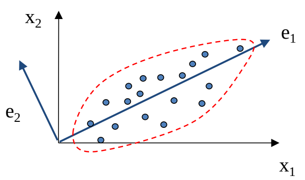
- Why data reduction?
  - Mining cost often increases rapidly with data size and number of attributes
- Goal: reduce data size, but produce (almost) the same results
- Data reduction strategies
  - Dimensionality reduction
  - Data Compression
  - Numerosity reduction
  - Discretization

Dimensionality Reduction: Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of attributes such that the mining result is still as good as (or even better than) when using all attributes
- Heuristic methods (due to exponential number of choices):
  - Select independently based on some test
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Eliminate attributes that some trusted method did not use, e.g., a decision tree

Principal Component Analysis

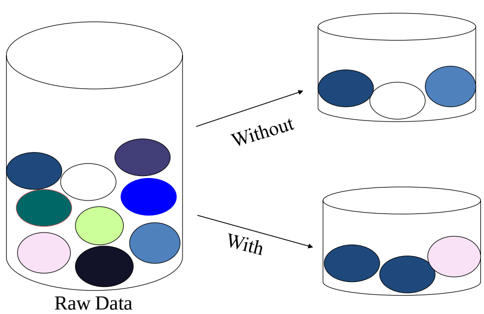
- Find projection that captures largest amount of variation in the data
  - Space defined by eigenvectors of the covariance matrix
- Compression: use only first k eigenvectors



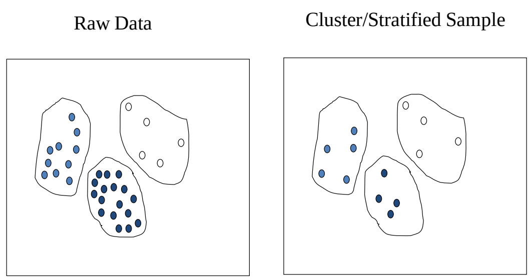
Data Reduction Method: Sampling

- Select a small subset of a given data set
- Reduces mining cost
  - Mining cost usually is super-linear in data size
  - Often makes difference between in-memory processing and need for expensive I/O
- Choose a representative subset of the data
  - Simple random sampling may have poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling
    - Approximate the percentage of each class (or sub-population of interest) in the overall database
    - Used in conjunction with skewed data

Sampling with or without Replacement



Sampling: Cluster or Stratified Sampling



Data Reduction: Discretization

- Applied to continuous attributes
- Reduces domain size
- Makes the attribute discrete and hence enables use of techniques that only accept categorical attributes
- Approach:
  - Divide the range of the attribute into intervals
  - Interval labels replace the original data

Data Preprocessing Overview

- Descriptive data summarization
- Data cleaning
- Data integration
- Data transformation
- Summary

Summary

- Data preparation is a big issue for data mining
- Descriptive data summarization is used to understand data properties
- Data preparation includes
  - Data cleaning and integration
  - Data reduction and feature selection
  - Discretization
- Many techniques and commercial tools, but still major challenge and active research area