

Master of Technology in Knowledge/Software Engineering

KE5107: Data Mining Methodology and Methods

Data Preparation and Transformation

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
E-mail: zhenzhen@nus.edu.sg

© 2016 NUS. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.

Module Objectives

- To use different methods to prepare and transform the data for data mining

Agenda

- The need to prepare data
- Major tasks in data preprocessing
 - Data cleaning
 - Data integration
 - Data transformation
 - Data reduction

Data Preparation and Transformation

- Data Cleaning and Preparation = Data Preprocessing, a critical phase in data mining
- 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
 - e.g., Salary="-10"
 - **inconsistent**: containing discrepancies in codes or names
 - e.g., Age="42" Birthday="03/07/1988"

Quality of Data Matters

- No quality data, no quality results
 - e.g., duplicate or missing data may cause incorrect or even misleading statistics



Major Tasks in Data Preprocessing

1. Data cleaning

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

2. Data integration

- Integration of multiple databases

3. Data transformation

- Normalization and aggregation

4. Data reduction

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

1. Data Cleaning

- Data may not be perfectly collected, or collected with the right purpose.
- Many reasons exist for data to be dirty:
 - Data entry errors
 - Misplaced decimal points
 - Inherent error in counting or measuring devices
 - External factors, etc.
- Data exploration can discover anomalous patterns, leading to the questioning of data quality
 - E.g. categories with very low frequency counts → mistyping?
 - Name and addresses recorded in multiple ways in data integrated from multiple sources (can be up to 20~30 variations)
 - Missing data

Data Cleaning Tasks

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

Missing Values

- Common feature of any dataset
- Various reasons:
 - Information not available
 - Lost data
 - Purposefully left out with a reason
- Might be marked with a special value
 - E.g. “9999”, “1 Jan 1900”, “*”, “?”, “#”, “\$”, etc
- The presence of missing values in data can make problems for the modeling tools.

Handling Missing Values

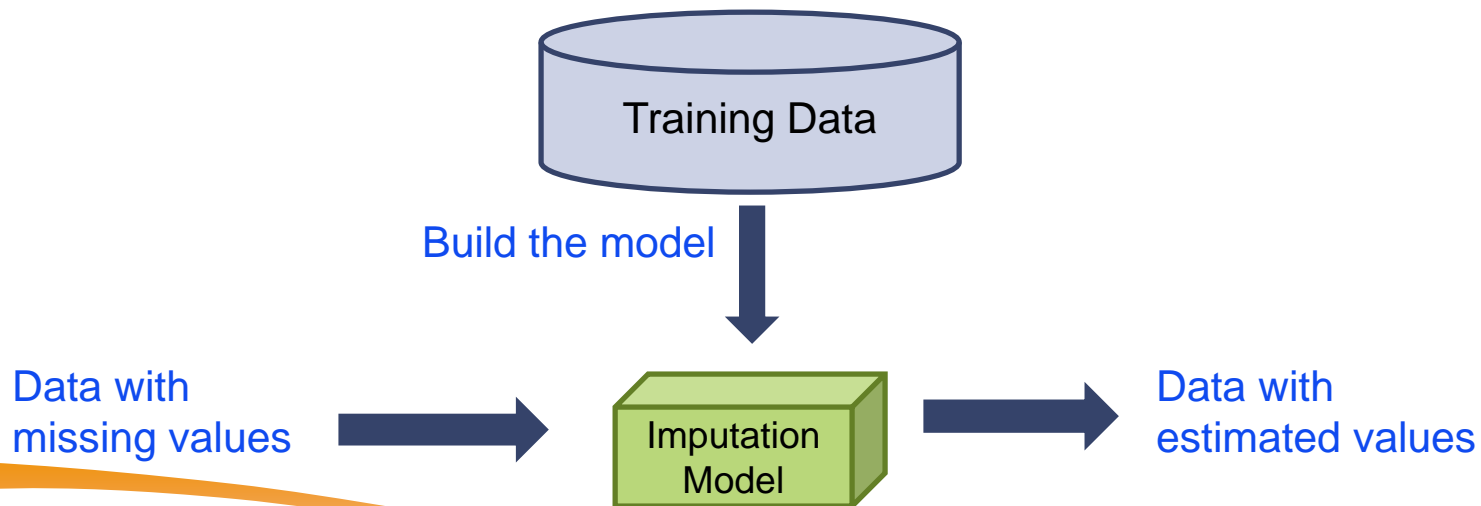
- Simply ignore the data sample with missing values
 - Throwing away data, especially poor when the percentage of missing values per attribute varies considerably
- Fill in the missing values – **data imputation**
 - Principle: Avoid adding bias and distortion to the data
 - Understand why the data is missing can help guide the imputation
 - E.g. for '*rainfall*' variable, a missing value may mean no rain recorded on that day → 0

Data Imputation

- Fill in the missing values automatically
 - a global constant : e.g., “unknown”
 - Straight forward approach
 - Modeling algorithms may mistakingly treat “unknown” as a concept
 - the attribute mean (or median, mode)
 - Simple and quick, been found empirically useful, though not always satisfactory
 - the attribute mean for all samples belonging to the same class
 - Better estimate than attribute mean

Data Imputation

- Train a prediction model (E.g. regression model, decision tree, k-Nearest Neighbor) to predict the **most probable** value
 - Use variables containing values to estimate the variable with missing values
 - Can produce good estimates
 - Need training data and additional modeling

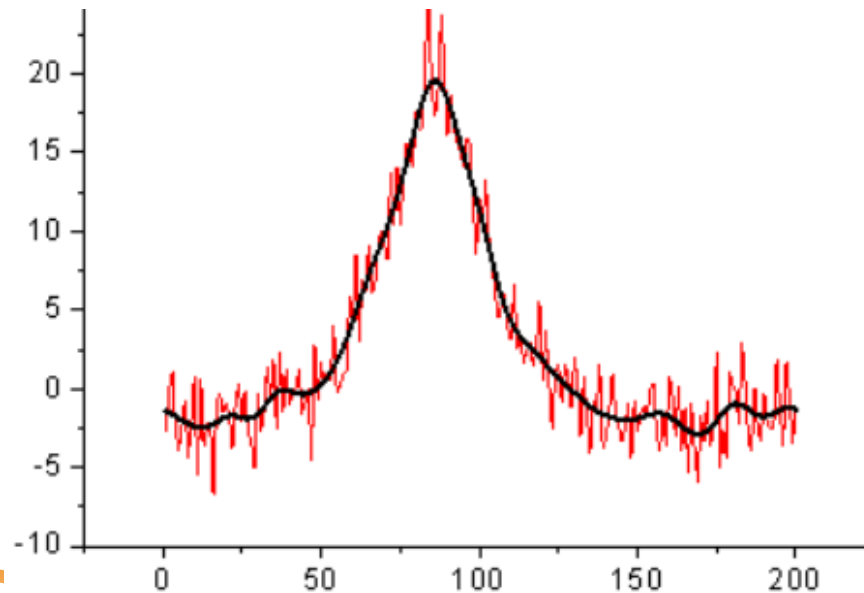


Noisy Data

- **Noise:** random error or variance in a measured variable
- Incorrect attribute values may have been entered due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention

Handling Noisy Data

- Binning method:
 - first sort data and partition into bins
 - then one can **smooth** the data by bin means, by bin median, by bin boundaries, etc.
 - Equal width binning, equal frequency binning , quantile binning, etc.



Handling Noisy Data

- 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)

Outliers

- Observations that “deviate so much from other observations as to arouse suspicion that it was generated by a different mechanism”. (Hawkins, 1980)
- Appearing at the maximum or minimum end of a variable, skewing or distorting the distribution
 - E.g. extreme weather conditions on a particular day, a very wealthy person financially very different from the rest of the population, etc.



Dealing with Outliers

- Can be rare, unusual, infrequent events we are interested in. They should be identified for further investigation.
 - E.g. frauds in income tax, insurance, banking, etc.
- Otherwise, outliers usually should be removed to avoid adversely affecting the modeling result (though some algorithms, like random forests and support vector machines can be robust to outliers)



2. Data Integration

- Data integration
 - combines data from multiple sources into a coherent store
- Detecting and resolving **data value conflicts**
 - for the same real world entity, attribute values from different sources may be different
 - possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- **Redundant data** occur often when integrating multiple databases
 - The same attribute may have different names in different databases
 - One attribute may be a “derived” attribute in another database, e.g., monthly revenue and annual revenue
- Redundant data may 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

3. Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
- Generalization: concept hierarchy climbing
- Log Transformation
- Square Root Transformation
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction
 - New attributes constructed from the given ones



Data Transformation

- Log Transformation
 - Makes a skewed attribute more symmetric
 - Reduces the magnitudes
 - Common bases 10, 2, e
- Square Root Transformation
 - More discrete values to normal form
- Categorical variable to indicator or dummy variables with value of 0 and 1 (some modelling algorithms include this step)

– E.g.

Obs.	Colour	Colour_Red	Colour_Green	Colour_Blue
1	Green	0	1	0
2	Blue	0	0	1
3	Blue	0	0	1
4	Red	1	0	0
5	Green	0	1	0
6	Red	1	0	0

Data Transformation: Normalization

- Reduces outlier distortion and enhances linear predictability
- Required before distance-measure-based clustering , to ensure all variables have approximately the same scale
 - E.g. variable *Age* vs *Income*: a distance of 10 “years” may be more significant than a distance of \$1000, yet \$1000 swamps 10 when they are added in calculating distance
- Normally re-center and rescale the data to be around zero, in the range from 0 to 1, etc.

Common Normalization Methods

- **min-max** normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- **z-score** normalization: centered around zero

$$v' = \frac{v - \text{mean}_A}{\text{stand_dev}_A}$$

- normalization by **decimal scaling**

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \text{Max}(|v'|) < 1$$

Attribute/Feature Construction

Product Holdings

1. ProdA + ProdC

2. ProdB + ProdC

3. ProdA + ProdD

4. ProdB + ProdD

...

Purchased Service

Y

N

N

Y



Proda

ProdB

ProdC

ProdD

Svc

1. 1

0

1

0

Y

2. 0

1

1

0

N

3. 1

0

0

1

N

4. 0

1

0

1

Y

...

4. Data Reduction

- Complex data analytics 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
 - Dimensionality reduction—reduce number of attributes
 - Numerosity reduction—fit data into models



Dimensionality Reduction

- Rational: not all variables contribute useful information
 - E.g. CRM data supplied by French telecom company Orange for 2009 KDD Cup has 15,000 variables. Are they all useful?
 - Irrelevant variables add noise to the data and increases the time and resources required for model building and scoring.
 - Noise increases the size of model and affect model accuracy
 - Groups of correlated variables (possibly measuring the same underlying feature)
 - Some algorithms, like nearest-neighbor learning, are susceptible to irrelevant features

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection)
 - selecting the most relevant attributes

315.62	316.71	317.72	318.29	318.18	316.54	314.7	313.84	313.18	315.98	315.9
316.43	317.58	319.02	320.03	319.19	318.18	315.9	314.16	313.9	316.19	316.9
316.93	318.54	319.48	320.57	319.48	318.57	316.9	314.8	313.9	317.01	317.6
317.94	319.59	320.63	321.6	319.59	319.59	317.9	316.25	315.2	316.69	317.7
318.74	319.66	321.39	322.4	319.74	319.74	318.7	317.2	316.2	317.12	318.2
319.57	320.69	322.1	323.1	319.57	320.44	319.5	317.79	316.9	317.79	318.9
319.8	320.69	322.1	323.1	319.8	320.69	322.1	317.8	316.8	317.8	318.8
320.62	321.5	322.39	323.87	320.62	322.39	323.87	318.64	317.1	319.79	321.08
322.06	323.04	324.42	325.8	322.06	323.04	324.42	319.31	318.1	320.72	321.96
322.57	323.89	325.02	326.1	322.57	323.89	325.02	320.41	319.1	321.31	322.84
324	325.4	326.64	327.8	324	325.4	326.64	321.1	320.1	321.98	323.13
325.03	326.47	327.8	329.1	325.03	326.47	327.8	321.8	320.8	322.8	324.0
325.17	327.18	327.78	329.2	325.17	327.18	327.78	322.36	321.4	323.4	324.6
325.77	327.75	328.05	329.42	325.77	327.75	328.05	323.03	322.0	324.03	325.17
326.58	328.4	328.7	330.1	326.58	328.4	328.7	323.71	322.7	324.7	325.8
327.35	329.1	329.4	330.8	327.35	329.1	329.4	324.3	323.4	325.4	326.5
328.15	329.8	330.1	331.5	328.15	329.8	330.1	325.0	324.1	326.1	327.2
329.35	330.8	331.1	332.5	329.35	330.8	331.1	325.7	324.8	326.8	327.9
330.4	331.1	332.04	333.1	330.4	331.1	332.04	326.4	325.5	327.5	328.6
331.78	332.4	333.8	334.8	331.78	332.4	333.8	327.1	326.2	328.2	329.3
332.93	334.1	335.7	336.7	332.93	334.1	335.7	327.8	326.9	328.9	330.0
334.23	336.4	338.0	339.0	334.23	336.4	338.0	328.5	327.6	329.6	330.7
335.23	337.73	339.3	340.3	335.23	337.73	339.3	329.2	328.3	330.3	331.4
336.31	340.08	340.7	341.7	336.31	340.08	340.7	329.9	329.0	330.9	332.0
337.23	341.38	342.51	343.5	337.23	341.38	342.51	330.6	329.7	331.6	332.7
340.75	344.7	345.7	346.7	340.75	344.7	345.7	331.3	330.4	332.3	333.4
341.37	346.94	347.1	348.1	341.37	346.94	347.1	332.0	331.1	333.0	334.1
343.7	348.4	348.7	349.7	343.7	348.4	348.7	332.7	331.8	333.7	334.8
344.97	349.73	349.9	350.9	344.97	349.73	349.9	333.4	332.5	334.4	335.5
346.3	350.9	351.2	352.2	346.3	350.9	351.2	334.1	333.2	335.1	336.2
350.43	354.7	355.7	356.7	350.43	354.7	355.7	334.8	333.9	335.8	336.9
352.76	357.1	358.1	359.1	352.76	357.1	358.1	335.5	334.6	336.5	337.6
353.66	358.3	359.3	360.3	353.66	358.3	359.3	336.2	335.3	337.2	338.3
355.7	360.5	361.5	362.5	355.7	360.5	361.5	336.9	336.0	337.9	339.0
357.97	362.7	363.7	364.7	357.97	362.7	363.7	337.6	336.7	338.6	339.7
360.25	364.9	365.9	366.9	360.25	364.9	365.9	338.3	337.4	339.3	340.4
363.18	367.1	368.1	369.1	363.18	367.1	368.1	339.0	338.1	340.0	341.1



315.62	316.38	316.71	317.72	318.29	318.18	316.54
316.43	317.58	317.58	319.02	320.03	319.19	318.18
316.93	318.54	319.48	320.57	320.57	319.48	318.57
317.94	319.59	320.63	321.6	321.6	319.59	319.59
318.74	319.66	321.39	322.4	322.4	319.66	321.39
319.57	320.69	322.1	323.1	323.1	320.69	322.1
319.8	320.69	322.1	323.1	323.1	320.69	322.1
320.62	321.5	322.39	323.87	323.87	321.5	322.39
322.06	323.04	324.42	325.8	325.8	323.04	324.42
322.57	323.89	325.02	326.1	326.1	323.89	325.02
324	325.4	326.64	327.8	327.8	325.4	326.64
325.03	326.47	327.8	329.1	329.1	326.47	327.8
325.17	327.18	327.78	329.2	329.2	327.18	327.78
325.77	327.75	328.05	329.42	329.42	327.75	328.05
326.58	328.4	328.7	330.1	330.1	328.4	328.7
327.35	329.1	329.4	330.8	330.8	329.1	329.4
328.15	329.8	330.1	331.5	331.5	329.8	330.1
329.35	330.8	331.1	332.5	332.5	330.8	331.1
330.4	331.1	332.04	333.1	333.1	331.1	332.04
331.78	332.4	333.8	334.8	334.8	332.4	333.8
332.93	334.1	335.7	336.7	336.7	334.1	335.7
334.23	336.4	338.0	339.0	339.0	336.4	338.0
335.23	337.73	339.3	340.3	340.3	337.73	339.3
336.31	340.08	340.7	341.7	341.7	340.08	340.7
337.23	341.38	342.51	343.5	343.5	341.38	342.51
340.75	344.7	345.7	346.7	346.7	344.7	345.7
341.37	346.94	347.1	348.1	348.1	346.94	347.1
343.7	348.4	348.7	349.7	349.7	348.4	348.7
344.97	349.73	349.9	350.9	350.9	349.73	349.9
346.3	350.9	351.2	352.2	352.2	350.9	351.2
350.43	354.7	355.7	356.7	356.7	354.7	355.7
352.76	357.1	358.1	359.1	359.1	357.1	358.1
353.66	358.3	359.3	360.3	360.3	358.3	359.3
355.7	360.5	361.5	362.5	362.5	360.5	361.5
357.97	362.7	363.7	364.7	364.7	362.7	363.7
360.25	364.9	365.9	366.9	366.9	364.9	365.9
363.18	367.1	368.1	369.1	369.1	367.1	368.1

- Feature Extraction
 - combining attributes into a new reduced set of features

315.62	316.38	316.71	317.72	318.29	318.18	316.54	314.7	313.84	313.18	315.98	315.9
316.43	317.58	317.58	319.02	320.03	319.19	318.18	315.9	314.16	313.9	316.19	316.9
316.93	318.54	319.48	320.57	320.57	319.48	318.57	316.9	314.8	313.9	317.01	317.6
317.94	319.59	320.63	321.6	321.6	319.59	319.59	317.9	316.25	315.2	316.69	317.7
318.74	319.66	321.39	322.4	322.4	319.66	321.39	318.7	317.2	316.2	317.12	318.2
319.57	320.69	322.1	323.1	323.1	320.69	322.1	319.5	317.79	316.9	317.79	318.9
319.8	320.69	322.1	323.1	323.1	320.69	322.1	319.8	317.8	316.8	317.8	318.8
320.62	321.5	322.39	323.87	323.87	321.5	322.39	320.6	317.1	319.79	321.08	322.3
322.06	323.04	324.42	325.8	325.8	323.04	324.42	322.0	318.1	320.72	321.96	323.1
322.57	323.89	325.02	326.1	326.1	323.89	325.02	322.5	319.1	321.31	322.84	324.0
324	325.4	326.64	327.8	327.8	325.4	326.64	324.0	320.1	321.98	323.13	324.6
325.03	326.47	327.8	329.1	329.1	326.47	327.8	325.0	321.0	322.8	324.0	325.1
325.17	327.18	327.78	329.2	329.2	327.18	327.78	325.1	321.8	322.8	324.0	325.1
325.77	327.75	328.05	329.42	329.42	327.75	328.05	325.7	322.4	323.4	324.6	325.7
326.58	328.4	328.7	330.1	330.1	328.4	328.7	326.5	323.0	324.0	325.1	326.2
327.35	329.1	329.4	330.8	330.8	329.1	329.4	327.3	323.7	324.7	325.8	326.9
328.15	329.8	330.1	331.5	331.5	329.8	330.1	328.1	324.3	325.4	326.5	327.6
329.35	330.8	331.1	332.5	332.5	330.8	331.1	329.3	325.0	326.1	327.2	328.3
330.4	331.1	332.04	333.1	333.1	331.1	332.04	330.4	325.7	326.8	327.9	329.0
331.78	332.4	333.8	334.8	334.8	332.4	333.8	331.7	326.4	327.5	328.6	329.7
332.93	334.1	335.7	336.7	336.7	334.1	335.7	332.9	327.1	328.2	329.3	330.4
334.23	336.4	338.0	339.0	339.0	336.4	338.0	334.2	327.8	328.9	330.0	331.1
335.23	337.73	339.3	340.3	340.3	337.73	339.3	335.2	328.5	329.6	330.7	331.8
336.31	340.08	340.7	341.7	341.7	340.08	340.7	336.3	329.2	330.3	331.4	332.5
337.23	341.38	342.51	343.5	343.5	341.38	342.51	337.2	329.9	331.0	332.1	333.2
340.75	344.7	345.7	346.7	346.7	340.75	344.7	340.7	330.6	331.7	332.8	333.9
341.37	346.94	347.1	348.1	348.1	341.37	346.94	341.3	331.3	332.4	333.5	334.6
343.7	348.4	348.7	349.7	349.7	343.7	348.4	343.7	332.0	333.1	334.2	335.3
344.97	349.73	349.9	350.9	350.9	344.97	349.73	344.9	332.7	333.8	334.9	336.0
346.3	350.9	351.2	352.2	352.2	346.3	350.9	346.3	333.4	334.5	335.6	336.7
350.43	354.7	355.7	356.7	356.7	350.43	354.7	350.4	334.1	335.2	336.3	337.4
352.76	357.1	358.1	359.1	359.1	352.76	357.1	352.7	334.8	335.9	337.0	338.1
353.66	358.3	359.3	360.3	360.3	353.66	358.3	353.6	335.5	336.6	337.7	338.8
355.7	360.5	361.5	362.5	362.5	355.7	360.5	355.7	336.2	337.3	338.4	339.5
357.97	362.7	363.7	364.7	364.7	357.97	362.7	357.9	336.9	338.0	339.1	340.2
360.25	364.9	365.9	366.9	366.9	360.25	364.9	360.2	337.6	338.7	339.8	340.9
363.18	367.1	368.1	369.1	369.1	363.18	367.1	363.1	338.3	339.4	340.5	341.6

Feature 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
- Reduce number of features, easier to understand
- Need a good knowledge of the data set and relevance of variables to the problem

Feature Selection

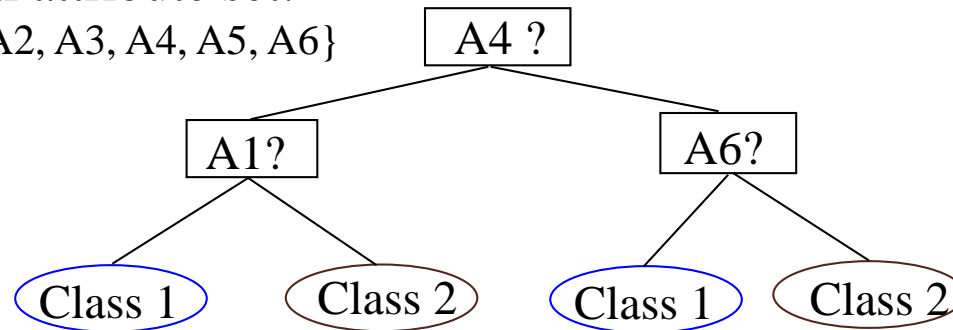
- Heuristic methods searching the attribute space (due to exponential # of choices):
 - Greedy search
 - Need evaluation measure (model accuracy)
 - **Step-wise forward selection**: start with a best feature and add to it the feature that together with the first provides the best result
 - **Step-wise backward elimination**: start with all features and remove the features one at a time for improved performance
 - Combining **forward selection** and **backward elimination**

Feature Selection

- Wrapper methods - a learning algorithm is wrapped into the selection procedure
 - decision-tree induction
 - E.g. use decision-tree on the full dataset to filter the variables to be used for nearest-neighbor algorithm
 - Algorithms that output coefficients of attributes (e.g. linear SVM)

Initial attribute set:

{A1, A2, A3, A4, A5, A6}



→ Reduced attribute set: {A1, A4, A6}

Feature Selection with Correlation Analysis

- Use correlation analysis to select a subset of attributes that individually correlate well with the target but have little inter-correlation
- Correlation analysis – given two attributes, such analysis measures how strongly one attribute implies the other, based on the available data.
 - Numeric data: correlation coefficient (such as *Pearson's coefficient*)
 - Categorical data: the χ^2 (chi-square) test
- Attributes found correlated with the target attribute are **important**.
- Attributes strongly correlated with each other are **redundant**.

Feature Extraction

- Also attribute reduction process by combining the original attributes
- Leading to a much smaller and richer set of attributes
- Methods exist which work well for linear between-variable relationships
 - Principle component analysis
 - Factor analysis

Principal Component Analysis

- Given N data vectors from k -dimensions, find $c \leq k$ orthogonal vectors that can be best used to represent data
 - The original data set is reduced to one consisting of N data vectors on c principal components (reduced dimensions)
 - A *component* is an artificially constructed variable that is fitted to all of the original variables in a dataset in such a way that it extracts the highest possible amount of variability
- Each data vector is a linear combination of the c principal component vectors
- Works for numeric data only
- Used when the number of dimensions is large
- Doesn't work well with nonlinear relationships

Factor Analysis

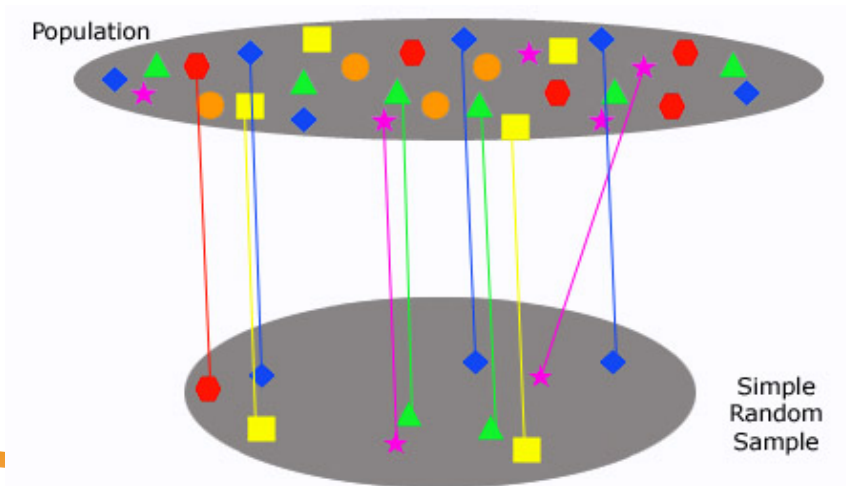
- Also a technique for reducing the complexity of high-dimensional data
- Statistical method to describe variability among observed, *correlated* variables in terms of a potentially lower number of unobserved variables called factors.
 - E.g. for psychiatric data, not possible to measure a factor of interest directly (such as intelligence), but possible to measure other quantities (like student exam scores, IQ test scores) that reflect the factor of interest.
 - Assumes that an unobserved variable is linearly related to two correlated variables
- Related to PCA, but differ in how the reduced dimensionality is constructed
 - Factor analysis uses regression modeling techniques where as PCA is a descriptive technique.
 - PCA explains variability; factor analysis explains correlation

Numerosity Reduction

- Data replaced or estimated by alternative, smaller data representations.
- Parametric methods
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data
- Non-parametric methods
 - Do not assume models
 - Major families: histograms, clustering, [sampling](#)

Sampling

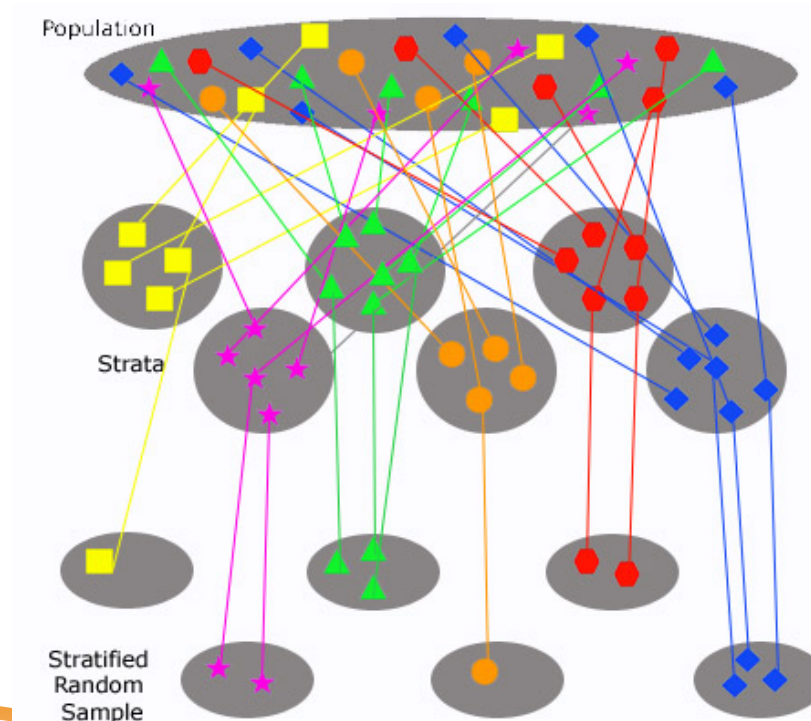
- Simple random sampling
 - without replacement
 - A sampled tuple doesn't participate in the next selection.
 - All tuples are equally likely to be sampled.
 - It avoids choosing any member of the population more than once.
 - with replacement
 - A sampled tuple is put back to the population for next selection



http://facultyweb.berry.edu/nmiller/classinfo/211/sampling_and_error_printout.htm

Sampling

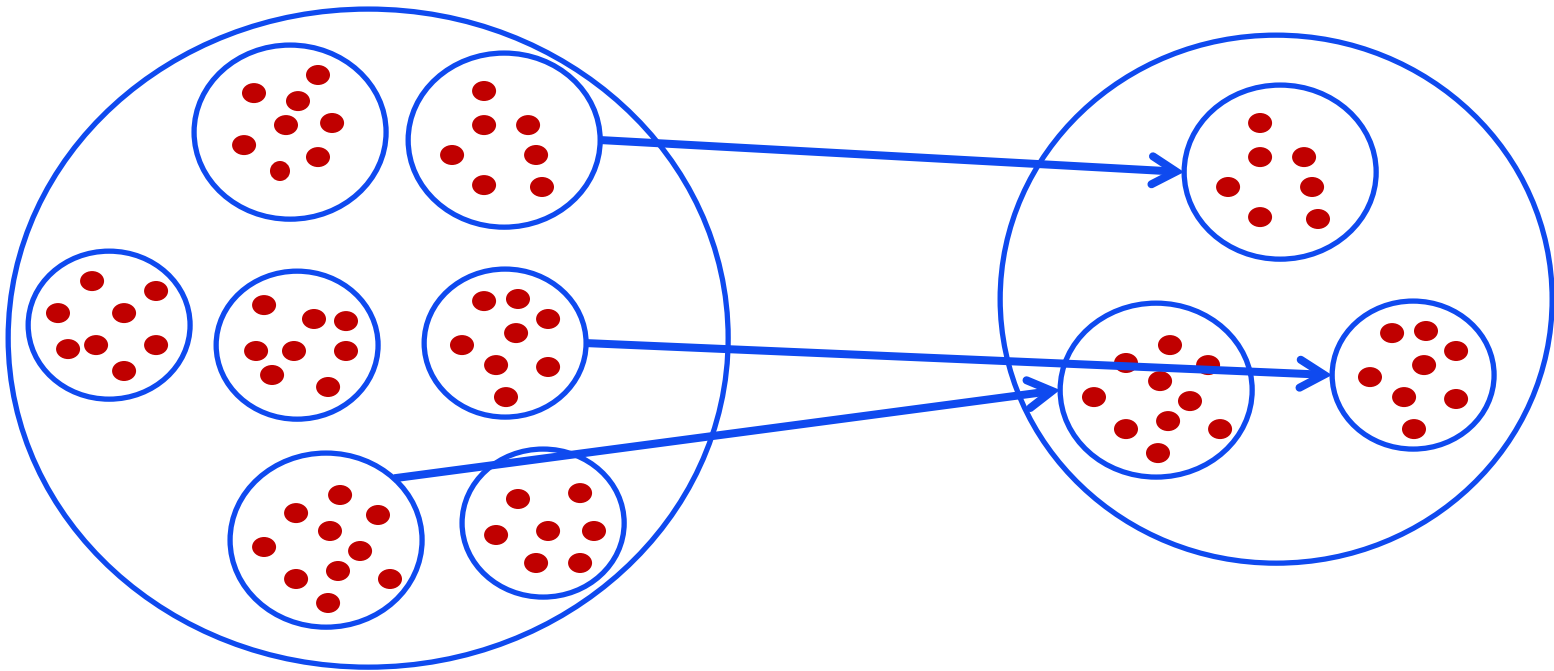
- Stratified sample
 - Divide data into mutually disjoint strata, and take simple random sample from each stratum
 - It helps ensure a representative sample, especially when the data are skewed



http://facultyweb.berry.edu/nmiller/classinfo/211/sampling_and_error_printout.htm

Sampling

- Cluster sampling
 - Group data into clusters, and take simple random sample of clusters instead of individuals



Summary

- Data needs to be prepared into a form suitable for further analysis and modeling.
- It's a tedious and time consuming process but essential for successful data mining.
- Data preparation improves the quality of data and consequently helps improve the quality of data mining results.

References

1. Jiawei Han and Micheline Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann Publishers, August 2000.
2. Tom Mitchell, *Machine learning*, McGraw Hill, 1997.
3. David Hand, Heikki Mannila, Padhraic Smyth, *Principles of data mining*, Cambridge, Mass. : MIT Press, c2001.
4. Graham Williams, *Data mining with Rattle and R*, Springer, 2011.