A Short Course in Data Mining

StatSoft®

STATISTICA

data analysis • data mining • quality control • web-based analytics

U.S. Headquarters: StatSoft, Inc. • 2300 E. 14th St. • Tulsa, OK 74104 • USA • (918) 749-1119 • Fax: (918) 749-2217 • info@statsoft.com • www.statsoft.com

Australia: StatSoft Pacific Pty Ltd.
Brazil: StatSoft South America
Bulgaria: StatSoft Bulgaria Ltd.
Czech Rep.: StatSoft Czech Rep. s.r.o.
China: StatSoft China
France:StatSoft France
Germany: StatSoft GmbH
Hungary: StatSoft Hungary Ltd.
India: StatSoft India Pvt. Ltd.
Israel: StatSoft Israel Ltd.
Italy: StatSoft Italia srl
Japan: StatSoft Japan Inc.
Korea: StatSoft Korea
Netherlands: StatSoft Benelux BV
Norway: StatSoft Norway AS
Poland: StatSoft Polska Sp. z o.o.
Portugal: StatSoft Ibérica Lda
Russia: StatSoft Russia
Spain: StatSoft Ibérica Lda
S. Africa: StatSoft S. Africa (Pty) Ltd.
Sweden: StatSoft Scandinavia AB
Taiwan: StatSoft Taiwan
UK: StatSoft Ltd.

Outline

Overview of Data Mining

- What is Data Mining?
- Models for Data Mining
- Steps in Data Mining
- Overview of Data Mining techniques
- Points to Remember

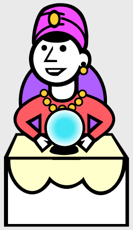
What is Data Mining?

- The need for Data Mining arises when expensive problems in business (manufacturing, engineering, etc.) have no obvious solutions
  - Optimizing a manufacturing process or a product formulation.
  - Detecting fraudulent transactions.
  - Assessing risk.
  - Segmenting customers.

A solution must be found.

- Pretend problem does not exist. Denial.

- Consult local psychic.



- Use data mining.

Note: We recommend this approach …

What is Data Mining?

- **Data mining** is an analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data.
- **Data mining** is a business process for maximizing the value of data collected by the business.
- **Data mining** is used to
  - Detect patterns in fraudulent transactions, insurance claims, etc.
  - Detect patterns in events and behaviors
  - Model customer buying patterns and behavior for cross-selling, up selling, and customer acquisition
  - Optimize product performance and manufacturing processes
  - <u>**Data mining can be utilized in any organization that needs to find patterns or relationships in their data, wherever the derived insights will deliver business value.**</u>

What is Data Mining?

The typical goals of data mining projects are:

- Identification of **groups, clusters, strata, or dimensions** in data that display no obvious structure,
- Identification of factors that are related to a particular outcome of interest (**root-cause** analysis)
- **Accurate prediction** of outcome variable(s) of interest (in the future, or in new customers, clients, applicants, etc.; this application is usually referred to as predictive data mining)

What is Data Mining?

- Data mining is a tool, not a magic wand.
- Data mining will not automatically discover solutions without guidance.
- Data mining will not sit inside of your database and send you an email when some interesting pattern is discovered.
- Data mining may find interesting patterns, but it does not tell you the value of such patterns.
- Data mining does not infer causality.
  - For example, it might be determined that males that have a certain income who exercise regularly are likely purchasers of a certain product, however, it does not mean that such factors cause them to purchase the product, only that the relationship exists.

Models for Data Mining

In the data mining literature, various “general frameworks” have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements.

- ***CRISP*** mid-1990s by a European consortium of companies to serve as a non-proprietary standard process model for data mining.

Business Understanding ↔ Data Understanding
↓
Data Preparation ↔ Modeling
↓
Evaluation
↓
Deployment

- ***DMAIC*** *Six Sigma* methodology - data-driven methodology for eliminating defects, waste, or quality control problems of all kinds.

Define → Measure → Analyze → Improve → Control

- ***SEMMA*** (SAS Institute) – focused more on technical aspects of data mining.

Sample → Explore → Modify → Model → Assess

Steps in Data Mining

Stage 0: Precise statement of the problem.

- Before opening a software package and running an analysis, the analyst must be clear as to what question he wants to answer. If you have not given a precise formulation of the problem you are trying to solve, then you are wasting time and money.

Stage 1: Initial exploration.

- This stage usually starts with data preparation that may involve the "cleaning" of the data (e.g., identification and removal of incorrectly coded data, etc.), data transformations, selecting subsets of records, and, in the case of data sets with large numbers of variables ("fields"), performing preliminary feature selection. Data description and visualization are key components of this stage (e.g. descriptive statistics, correlations, scatterplots, box plots, etc.).

Stage 2: Model building and validation.

- This stage involves considering various models and choosing the best one based on their predictive performance.

Stage 3: Deployment.

- When the goal of the data mining project is to predict or classify new cases (e.g., to predict the credit worthiness of individuals applying for loans), the third and final stage typically involves the application of the best model or models (determined in the previous stage) to generate predictions

Stage 1: Initial exploration

- **“Cleaning” of data**,
  - Identification and removal of incorrectly coded data, Male=Yes, Pregnant=Yes

- **Data transformations**,

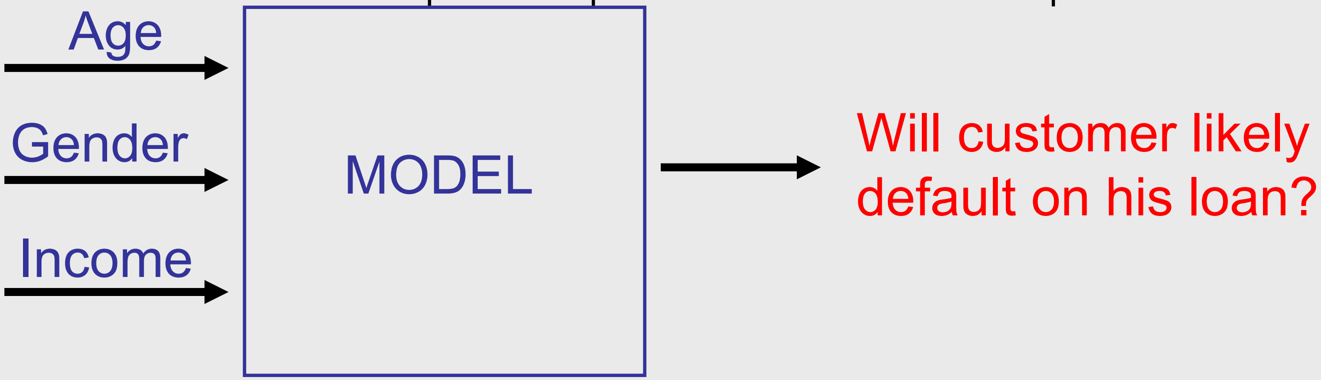
  Data may be skewed (that is, outliers in one direction or another
  may be present). Log transformation, Box-Cox transformation, etc.

- **Data reduction,** Selecting subsets of records, and, in the case of data sets with large numbers of variables (“fields”), performing preliminary feature selection.

- **Data description and visualization** are key components of this stage (e.g. descriptive statistics, correlations, scatterplots, box plots, brushing tools, etc.)
  - Data description allows you to get a snapshot of the important characteristics of the data (e.g. central tendency and dispersion).
  - Patterns are often easier to perceive visually than with lists and tables of numbers.

Stage 2: Model building and validation.

- Data Mining involves creating models of reality
- A model takes one or more inputs and produces one or more outputs



Age

Gender

Income

MODEL

Will customer likely default on his loan?

- A model can be "transparent", for example, a series of if/then statements where structure is easily discerned, or a model can be seen as a black box, for example, neural network, where the structure or the rules that govern the predictions are impossible to fully comprehend.

Stage 2: Model building and validation.

- A model is typically rated according to 2 aspects:
  - Accuracy
  - Understandability
- These aspects sometimes conflict with one another.
- Decision trees and linear regression models are less complicated and simpler than models such as neural networks, boosted trees, etc. and thus easier to understand, however, you might be giving up some predictive accuracy.
- Remember not to confuse the data mining model with reality (a road map is not a perfect representation of the road) but it can be used as a useful guide.

Generalization is the ability of a model to make accurate predictions when faced with data not drawn from the original training set (but drawn from the same source as the training set).

Stage 2: Model building and validation.

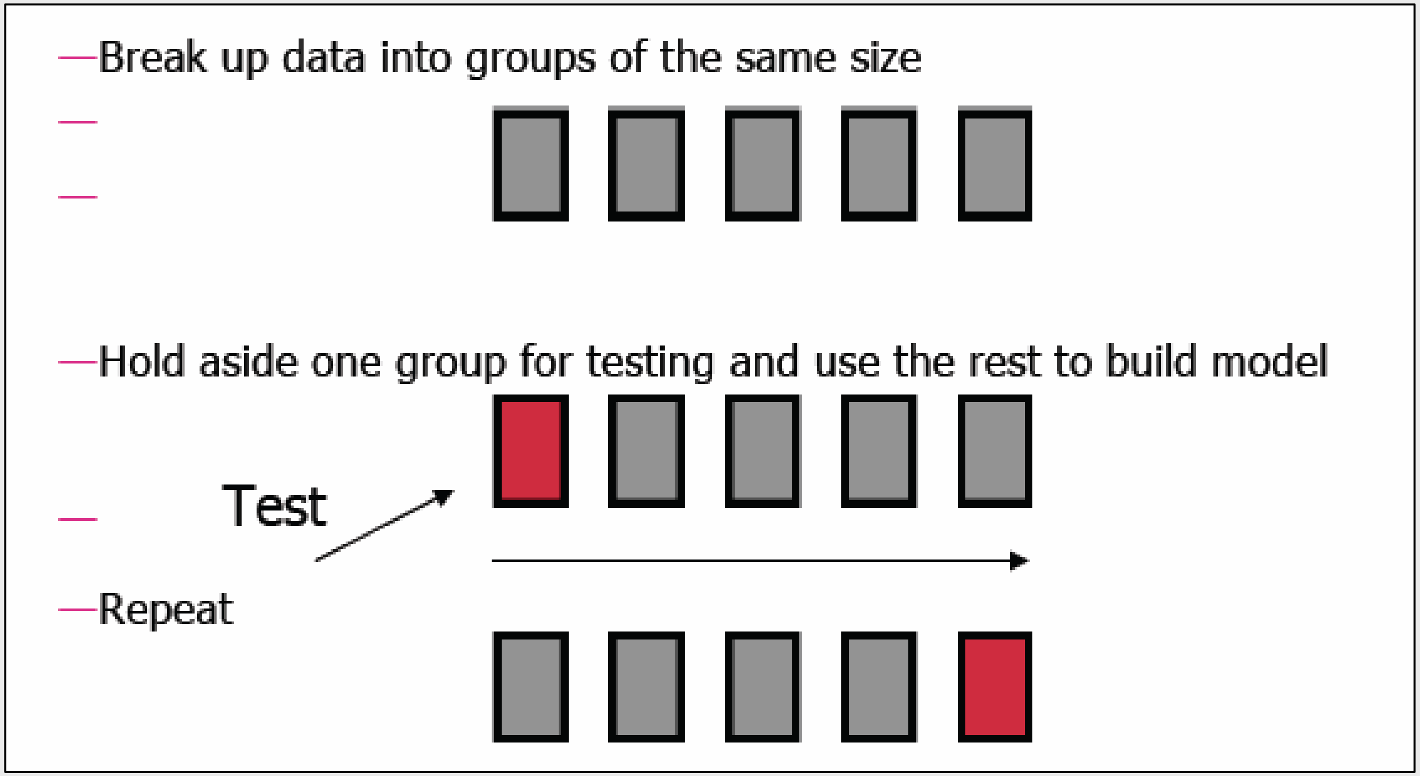
- Validation of the model requires that you train the model on one set of data and evaluate on another independent set of data.

- There are two main methods of validation
  - Split data into train/test datasets (75-25 split)
  - If you do not have enough data to have a holdout sample, then use v-fold cross validation.

In **v-fold cross-validation**, repeated (v) random samples are drawn from the data for the analysis, and the respective model or prediction method, etc. is then applied to compute predicted values, classifications, etc. Typically, summary indices of the accuracy of the prediction are computed over the v replications; thus, this technique allows the analyst to evaluate the overall accuracy of the respective prediction model or method in repeatedly drawn random samples.

Stage 2: Model building and validation.

- ➢ If you do not have enough data to have a holdout sample, then use v-fold cross validation.

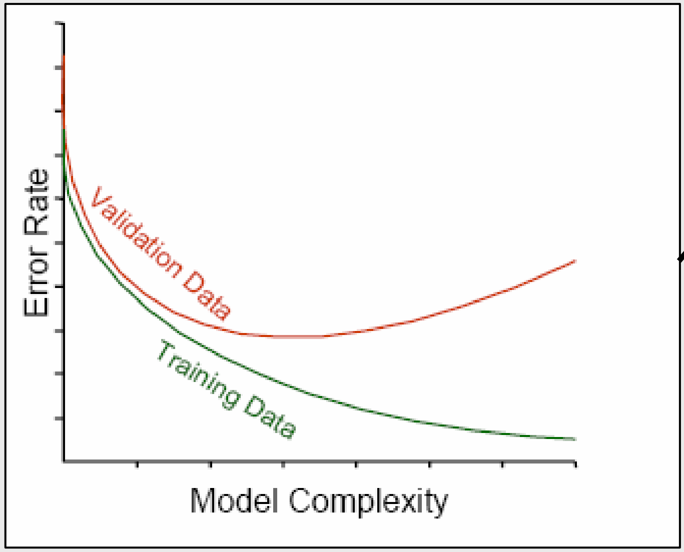


—Break up data into groups of the same size

—Hold aside one group for testing and use the rest to build model

—Repeat

Stage 2: Model building and validation.



In general, the term **overfitting** refers to the condition where a predictive model (e.g., for predictive data mining) is so "specific" that it reproduces various idiosyncrasies (random "noise" variation) of the particular data from which the parameters of the model were estimated; as a result, such models often may not yield accurate predictions for new observations (e.g., during deployment of a predictive data mining project). Often, various techniques such as cross-validation and v-fold cross-validation are applied to avoid **overfitting**.

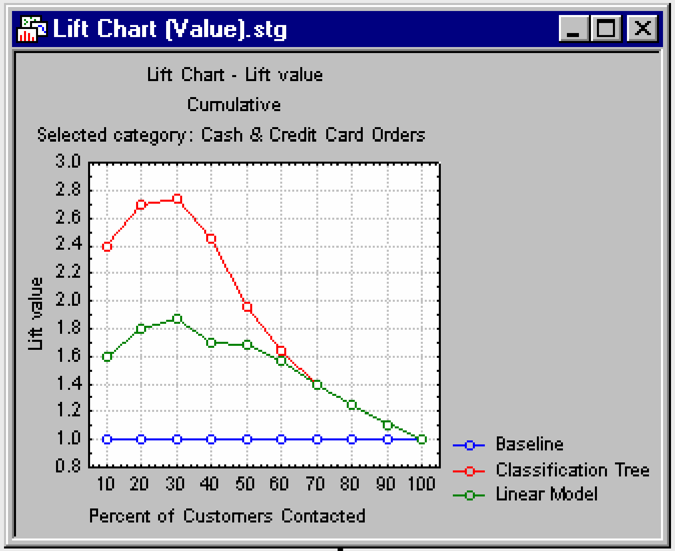
- How predictive is the model?
  - Compute error sums of squares (regression) or confusion matrix (classification)
- Error on training data is not a good indicator of performance on future data
  - The new data will probably not be exactly the same as the training data!
- **Overfitting:** Fitting the training data too precisely
  - Usually leads to poor results on new data

Stage 2: Model building and validation.

Model Validation Measures

- Possible validation measures
  - Classification accuracy
  - Total cost/benefit – when different errors involve different costs
  - Lift and Gains curves
  - Error in Numeric predictions

- Error rate
  - Proportion of errors made over the whole set of instances
  - Training set error rate: is way too optimistic!
  - You can find patterns even in random data



The **lift chart** provides a visual summary of the usefulness of the information provided by one or more statistical models for predicting a binomial (categorical) outcome variable (dependent variable); for multinomial (multiple-category) outcome variables, lift charts can be computed for each category. Specifically, the chart summarizes the utility that one may expect by using the respective predictive models compared to using baseline information only.

Stage 3: Deployment.

- A model is built once, but can be used over and over again.

- Model should be easily deployable.

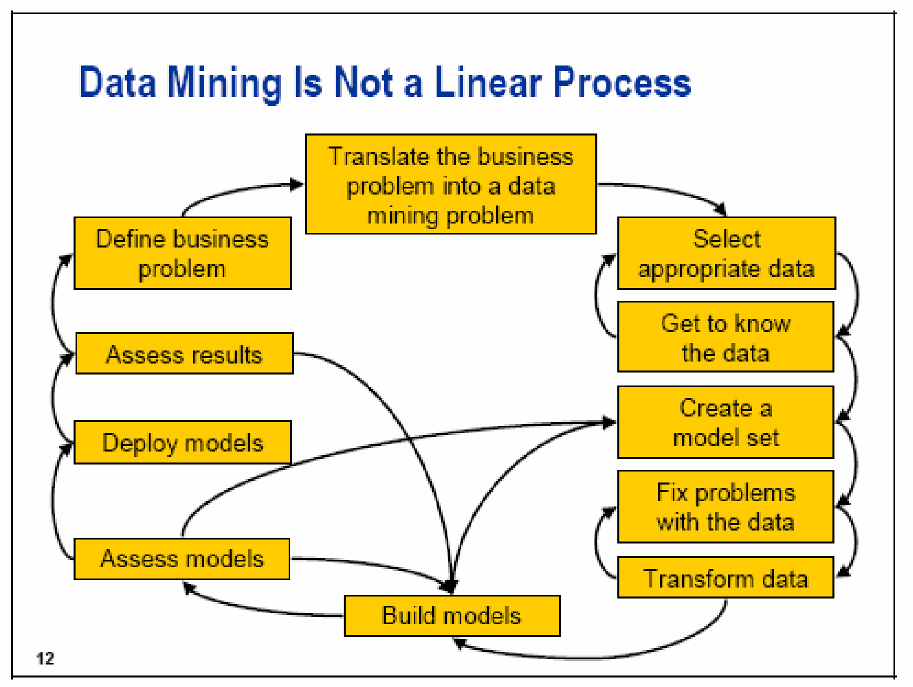
  - A linear regression is easily deployed. Simply gather the regression coefficients…
  - For example, if a new observed data vector comes in {x1, x2, x3}, then simply plug into linear equation to generate predicted value,

    Prediction = B0 + B1*X1 + B2*X2 + B3*X3

- A Classification and Regression Tree model is easily deployed: A series of If/Then/Else statements …

Steps in Data Mining

Stage 0: Precise statement of the problem.
Stage 1: Initial exploration.
Stage 2: Model building and validation.
Stage 3: Deployment.



Data Mining Is Not a Linear Process

Overview of Data Mining techniques

- Supervised Learning
  - **Classification**: response is categorical
  - **Regression**: response is continuous
  - **Time Series**: dealing with observations across time
  - **Optimization**: minimize or maximize some characteristic
- Unsupervised Learning
  - **Principal Component Analysis**: feature reduction
  - **Clustering**: grouping like objects together
  - **Association and Link Analysis**: descriptive approach to exploring data that helps identify relationships among values in a database, e.g. Market basket analysis, those customers that buy hammers also buy nails. Examine conditional probabilities.

Supervised Learning: A category of data mining methods that use a set of labeled training examples (e.g., each example consists of a set of values on predictors and outcomes) to fit a model that later can be used for deployment.
Unsupervised Learning: A data mining method based on training data where the outcomes are not provided.

Overview of Data Mining Techniques

- The view from 20,000 feet above.

  The next few slides will cover the commonly used data mining techniques below at a high level, focusing on the big picture so that you can see how each technique fits into the overall landscape.

- Descriptive Statistics
- Linear and Logistic Regression
- Analysis of Variance (ANOVA)
- Discriminant Analysis
- Decision Trees
- Clustering Techniques (K Means & EM)
- Neural Networks
- Association and Link Analysis
- MSPC (Multivariate Statistical Process Control)

Disadvantages of nonparametric models

- Some data mining algorithms need a lot of data…

  Curse of dimensionality is a term coined by Richard Bellman applied to the problem caused by the rapid increase in volume associated with adding extra dimensions to a (mathematical) space.

  Leo Breiman gives as an example the fact that 100 observations cover the one-dimensional unit interval [0,1] on the real line quite well. One could draw a histogram of the results, and draw inferences. If one now considers the corresponding 10-dimensional unit hypersquare, 100 observations are now isolated points in a vast empty space. To get similar coverage to the one-dimensional space would now require 10^{20} observations, which is at least a massive undertaking and may well be impractical.

The term **curse of dimensionality** (Bellman, 1961, Bishop, 1995) generally refers to the difficulties involved in fitting models in many dimensions. As the dimensionality of the input data space (i.e., the number of predictors) increases, it becomes exponentially more difficult to find global optima for the models. Hence, it is simply a practical necessity to pre-screen and preselect from among a large set of input (predictor) variables those that are of likely utility for predicting the outcomes of interest.

The curse of dimensionality is a significant obstacle in machine learning problems that involve learning from few data samples in a high-dimensional feature space.

See also :
http://en.wikipedia.org/wiki/Curse_of_dimensionality

Descriptive Statistics, etc.

- Typically there is so much data in a database that it first must be summarized in order to begin to make any sense of it. The first step in Data Mining is describing and summarizing the data.
- Two main types of statistics commonly used to characterize a distribution of data are:
  - **Measures of Central Tendency**

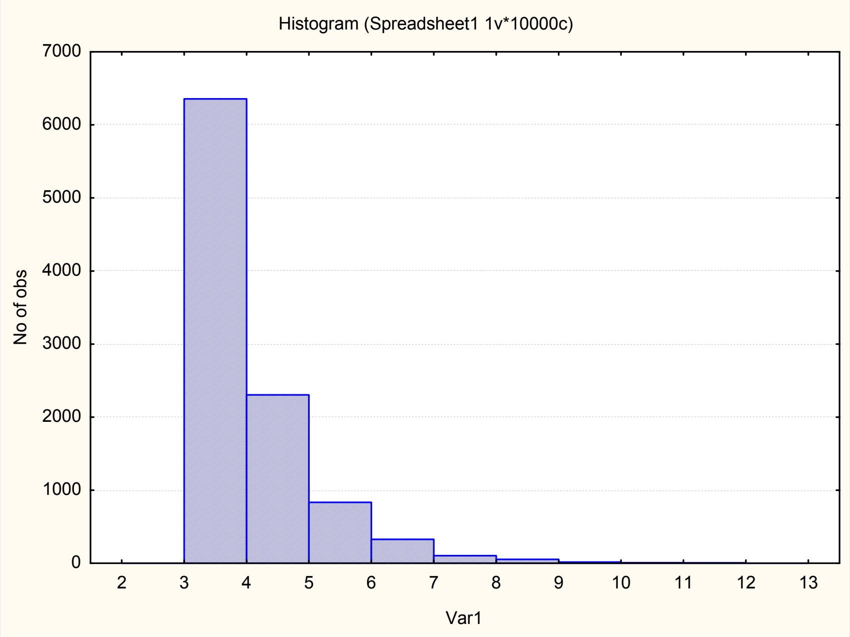
    Mean, Median, Mode
  - **Measures of Dispersion**

    Standard Deviation, Variance
- Visualization of the data is crucial. Patterns that can be seen by the eye leave a much stronger imprint than a table of numbers or statistics.

Descriptive Statistics, etc.

A histogram is a simple yet effective way of summarizing information in a column or variable.

We can quickly determine the range of the variable (min and max), the mean, median, mode, and variance.



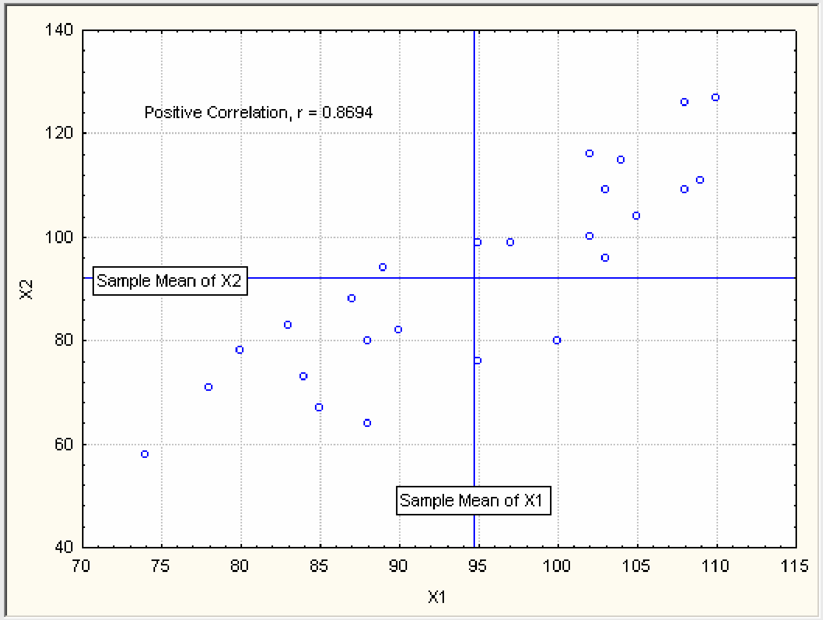
Variable	Descriptive Statistics (Spreadsheet1)						
	Valid N	Mean	Median	Minimum	Maximum	Variance	Std.Dev.
Var1	10000	3.992616	3.677125	3.000038	11.46572	0.995422	0.997708

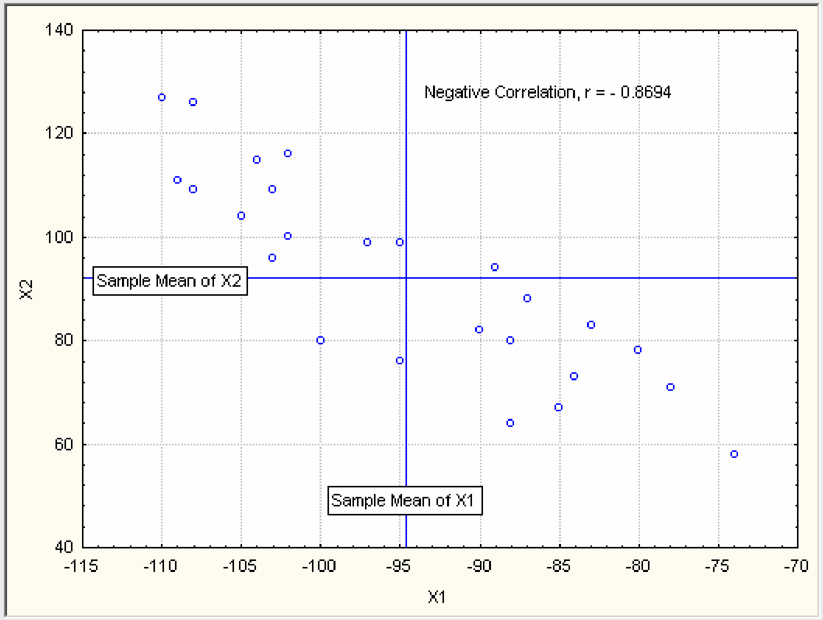
Descriptive Statistics, etc.

$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Positive correlation

Negative correlation





Linear and Logistic Regression

- Regression analysis is a statistical methodology that utilizes the relationship between two or more quantitative variables, so that one variable can be predicted from the others.

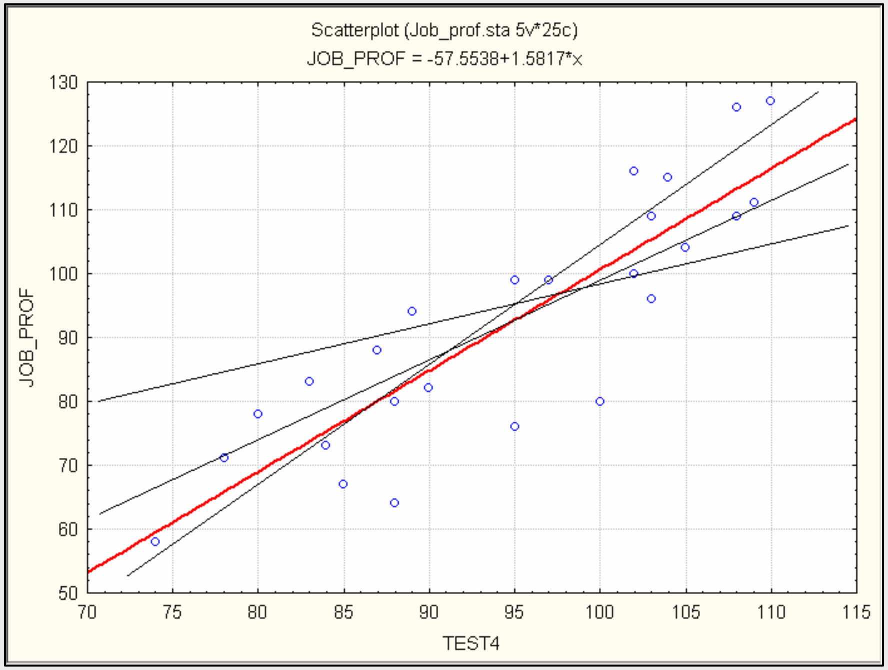
- We shall first consider linear regression. This is the case where the response variable is continuous. In logistic regression the response is dichotomous.

- Examples include:
  - Sales of a product can be predicted utilizing the relationship between sales and advertising expenditures.
  - The performance of an employee on a job can be predicted by utilizing the relationship between performance and a battery of aptitude tests.
  - The size of vocabulary of a child can be predicted by utilizing the relationship between size of vocabulary and age of child and amount of education of the parents.

Linear and Logistic Regression

The simplest form of regression contains one predictor and one response. Y = $\beta_0 + \beta_1 X$

The slope and intercept terms are found such that sum of squared deviations from the line are minimized. This is the principle of **least squares.**



Linear and Logistic Regression

- Personnel professionals customarily use multiple regression procedures to determine equitable compensation. The personnel analyst then usually conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics for different positions. This information can be used in a multiple regression analysis to build a regression equation of the form:

  Salary = .5 *(Amount of Responsibility) + .8 * (Number of People To Supervise)

- What if the pattern in the data is NOT linear?
  - More predictors can be used
  - Transformations
  - Interactions and higher order polynomial terms can be added (hence linear regression does not mean linear in the predictors, but rather linear in the parameters), that is, we can easily bend the line into curves that are nonlinear.
  - We can really develop complicated models using this approach, however, this takes expertise on the part of the modeler in both the domain of application as well as with the methodology. Techniques that we will look at later can do a lot of the “grunt” work for us….

Linear and Logistic Regression

- What if the response is dichotomous?

  - We might use linear regression to predict the 0/1 response. However, there are problems such as:

    - If you use linear regression, the predicted values will become greater than one and less than zero if you move far enough on the X-axis. Such values are theoretically inadmissible.
    - One of the assumptions of regression is that the variance of Y is constant across values of X (homoscedasticity). This cannot be the case with a binary variable, because the variance is PQ. When 50 percent of the people are 1s, then the variance is .25, its maximum value. As we move to more extreme values, the variance decreases. When P=.10, the variance is .1*.9 = .09, so as P approaches 1 or zero, the variance approaches zero.
    - The significance testing of the *b* weights rest upon the assumption that errors of prediction (Y-Y') are normally distributed. Because Y only takes the values 0 and 1, this assumption is pretty hard to justify, even approximately. Therefore, the tests of the regression weights are suspect if you use linear regression with a binary DV.

Linear and Logistic Regression

- There are a variety of regression applications in which the response variables only has two qualitative outcomes (male/female, default/no default, success/failure).
  - In a study of labor force participation of wives, as a function of age of wife, number of children, and husband’s income, the response variable was defined to have two outcomes: wife in labor force, wife not in labor force.
  - In a longitudinal study of coronary heart disease as a function of age, gender, smoking history, cholesterol level, and blood pressure, the response variable was defined to have to two possible outcomes: person developed heart disease, person did NOT develop heart disease during the study.

ANOVA

- Analysis of Variance (ANOVA) allows you to test differences among means.
- The explanatory variables in an ANOVA are typically qualitative (gender, geographic location, plant shift, etc.)
- If predictor variables are quantitative, then no assumption is made about the nature of the regression function between response and predictors.

- Some examples include:
  - An experiment to study the effects of five different brands of gasoline on automobile operating efficiency (mpg).
  - An experiment to assess the effects of different amounts of a particular psychedelic drug on manual dexterity.

Discriminant Analysis

- Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups.
- For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide

  (1) to go to college,

  (2) to attend a trade or professional school, or

  (3) to seek no further training or education.

  For that purpose the researcher could collect data on numerous variables prior to students' graduation. After graduation, most students will naturally fall into one of the three categories. Discriminant Analysis could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.
- A medical researcher may record different variables relating to patients' backgrounds in order to learn which variables best predict whether a patient is likely

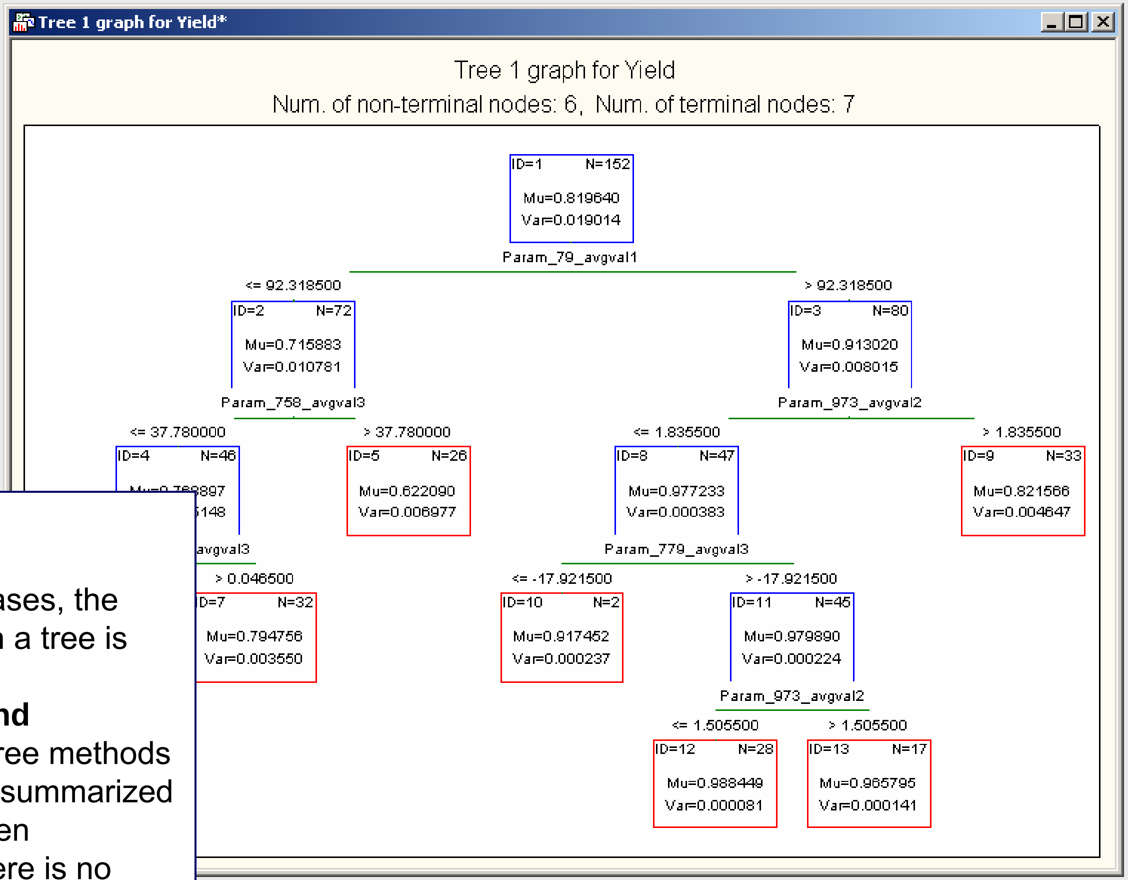
  (1.) to recover completely (group 1),

  (2.) partially (group 2), or

  (3.) not at all (group 3).

Decision Trees

- Decision trees are predictive models that can be viewed as a tree.
- Models can be made to predict categorical or continuous responses.
- A decision tree is nothing more than a sequence of if/then statements.


Tree 1 graph for Yield*
Tree 1 graph for Yield
Num. of non-terminal nodes: 6, Num. of terminal nodes: 7
ID=1 N=152 Mu=0.819640 Var=0.019014
Param_79_avgval1
<= 92.318500
> 92.318500
ID=2 N=72 Mu=0.715883 Var=0.010781
Param_758_avgval3
ID=3 N=80 Mu=0.913020 Var=0.008015
Param_973_avgval2
<= 37.780000
> 37.780000
ID=4 N=46
ID=5 N=26 Mu=0.622090 Var=0.006977
<= 1.835500
> 1.835500
ID=8 N=47 Mu=0.977233 Var=0.000383
ID=9 N=33 Mu=0.821566 Var=0.004647
avgval3
> 0.046500
D=7 N=32 Mu=0.794756 Var=0.003550
Param_779_avgval3
<= -17.921500
> -17.921500
ID=10 N=2 Mu=0.917452 Var=0.000237
ID=11 N=45 Mu=0.979890 Var=0.000224
Param_973_avgval2
<= 1.505500
> 1.505500
ID=12 N=28 Mu=0.988449 Var=0.000081
ID=13 N=17 Mu=0.965795 Var=0.000141

Some Advantages of Tree Models

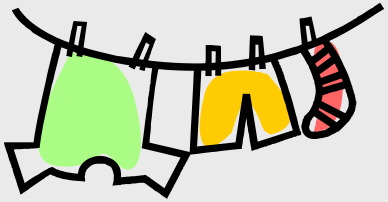
Easy to Interpret Results. In most cases, the interpretation of results summarized in a tree is very simple.
Tree methods are Nonparametric and Nonlinear. The final results of using tree methods for classification or regression can be summarized in a series of (usually few) logical if-then conditions (tree nodes). Therefore, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear, follow some specific non-linear link function, or that they are even monotonic in nature.

Clustering

- Clustering is the method in which like records are grouped together.

- A simple example of clustering would be the clustering that one does when doing laundry – grouping the permanent press, dry cleaning, whites, and brightly colored clothes.

- This is straightforward except for the white shirt with red stripes…where does this go?



Cluster Analysis: Marketing Application

A typical example application is a marketing research study where a number of consumer-behavior related variables are measured for a large sample of respondents; the purpose of the study is to detect "market segments," i.e., groups of respondents that are somehow more similar to each other (to all other members of the same cluster) when compared to respondents that "belong to" other clusters. In addition to identifying such clusters, it is usually equally of interest to determine how those clusters are different, i.e., the specific variables or dimensions on which the members in different clusters will vary, and how.

Clustering

Clustering techniques used to…

- Identifying characteristics of people belonging to different clusters (income, age, marital status, etc.).

Possible application of results…

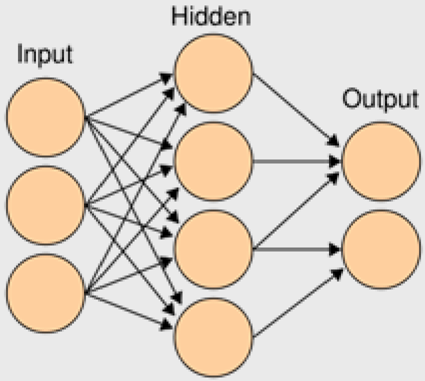
- Develop special marketing campaigns, services, or recommendations to particular type of stores based on characteristics.
- Arrange stores according to the taste of different shopper groups.
- Enhance the overall attractiveness and quality of the shopping experience….

K Means Clustering: The classic k-Means algorithm was popularized and refined by Hartigan (1975; see also Hartigan and Wong, 1978). The basic operation of that algorithm is relatively simple: Given a fixed number of (desired or hypothesized) k clusters, assign observations to those clusters so that the means across clusters (for all variables) are as different from each other as possible.

EM Clustering: The EM algorithm for clustering is described in detail in Witten and Frank (2001). The goal of EM clustering is to estimate the means and standard deviations for each cluster, so as to maximize the likelihood of the observed data (distribution). Put another way, the EM algorithm attempts to approximate the observed distributions of values based on mixtures of different distributions in different clusters.

Neural Networks

Like most statistical models neural networks are capable of performing three major tasks including *regression*, *classification*. Regression tasks are concerned with relating a number of input variables *x* with set of continuous *outcomes t* (target variables). By contrast, classification tasks assign class memberships to a categorical target variable given a set of input values. In the next section we will consider regression in more details.

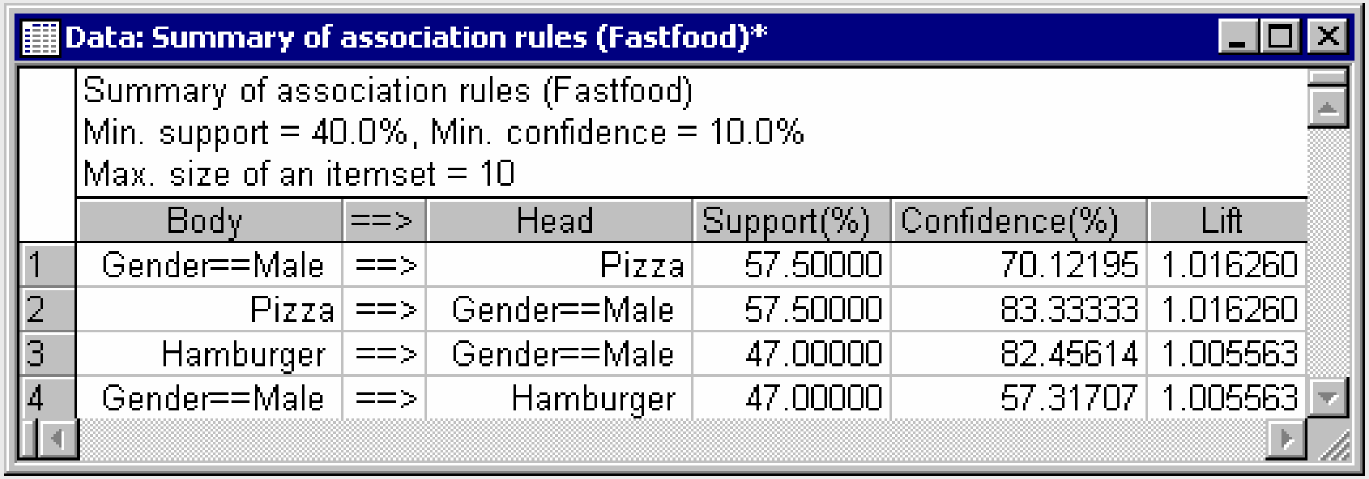


Association and Link Analysis

- Find items in a database that occurs together (***Association Rules***)
- An association is an expression of form:

Body Head (Support, Confidence)

If buys (x," flashlight") buys (x," batteries") (250, 89%)

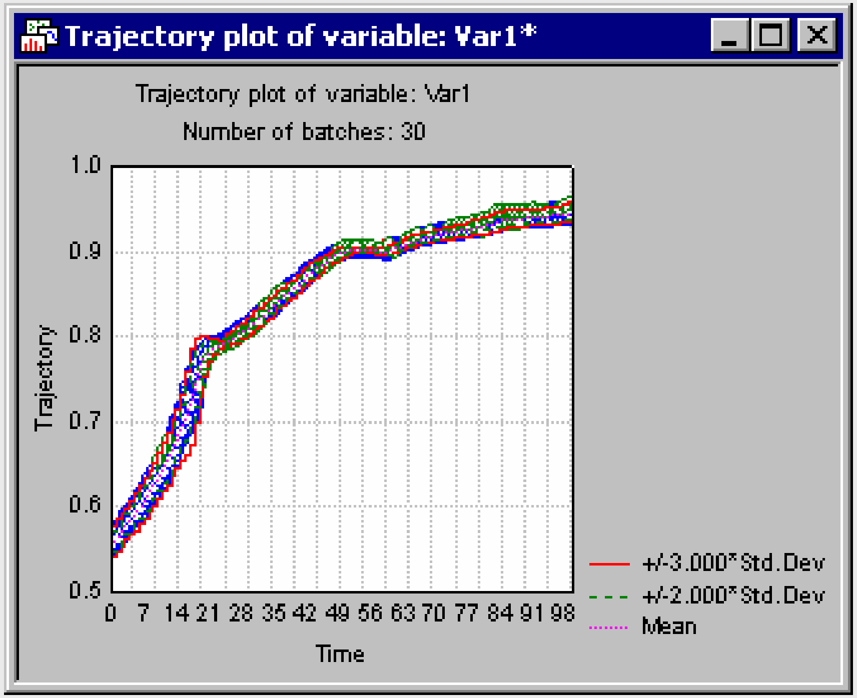


Summary of association rules (Fastfood)
Min. support = 40.0%, Min. confidence = 10.0%
Max. size of an itemset = 10

	Body	==>	Head	Support(%)	Confidence(%)	Lift
1	Gender==Male	==>	Pizza	57.50000	70.12195	1.016260
2	Pizza	==>	Gender==Male	57.50000	83.33333	1.016260
3	Hamburger	==>	Gender==Male	47.00000	82.45614	1.005563
4	Gender==Male	==>	Hamburger	47.00000	57.31707	1.005563

MSPC

- Built upon the capabilities of *PCA* and *PLS* techniques, ***MSPC*** is a selection of methods particularly designed for process monitoring and quality control in industrial batch processing.
- Batch processes are of considerable importance in making products with the desired specifications and standards in many sectors of the industry, such as polymers, paint, fertilizers pharmaceuticals/biopharm, cement, petroleum products, perfumes, and semiconductors.
- The objectives of batch processing are related to profitability achieved by reducing product variability as well as increasing quality.
- From a quality point of view, batch processes can be divided into normal and abnormal batches. Generally speaking, a normal batch leads to a product with the desired specifications and standards. This is in contrast to abnormal batch runs where the end product is expected to have a poor quality.



- Another reasons for batch monitoring is related to regulatory and safety purposes. Often industrial productions are required to keep full track (i.e., history) of the batch process for presentation of evidence on good quality control practice. *MSPC* helps construct an effective engineering system that can be used to monitor the progression of a batch and predict the quality of the end product.

Points to Remember…

- Data mining is a tool, not a magic wand.
- Data mining will not automatically discover solutions without guidance.
- Predictive relationships found via data mining are not necessarily causes of an action or behavior.
- To ensure meaningful results, it’s vital that you understand your data.
- **Data mining’s central quest:** Find meaningful, effective patterns and avoid overfitting (finding random patterns by searching too many possibilities)