Chapter 1

STATISTICAL METHODS FOR DATA MINING

Yoav Benjamini
Department of Statistics, School of Mathematical Sciences, Sackler Faculty for Exact Sciences
Tel Aviv University
ybenja@post.tau.ac.il

Moshe Leshno
Faculty of Management and Sackler Faculty of Medicine
Tel Aviv University
leshnom@post.tau.ac.il

Abstract The aim of this chapter is to present the main statistical issues in Data mining (DM) and Knowledge Data Discovery (KDD) and to examine whether traditional statistics approach and methods substantially differ from the new trend of KDD and DM. We address and emphasize some central issues of statistics which are highly relevant to DM and have much to offer to DM.

Keywords: Statistics, Regression Models, False Discovery Rate (FDR), Model selection and False Discovery Rate (FDR)

1. Introduction

In the words of anonymous saying there are two problems in modern science: too many people using different terminology to solve the same problems and even more people using the same terminology to address completely different issues. This is particularly relevant to the relationship between traditional statistics and the new emerging field of knowledge data discovery (KDD) and data mining (DM). The explosive growth of interest and research in the domain of KDD and DM of recent

years is not surprising given the proliferation of low-cost computers and the requisite software, low-cost database technology (for collecting and storing data) and the ample data that has been and continues to be collected and organized in databases and on the web. Indeed, the implementation of KDD and DM in business and industrial organizations has increased dramatically, although their impact on these organizations is not clear. The aim of this chapter is mainly to present the main statistical issues in DM and KDD and to examine the role of traditional statistics approach and methods in the new trend of KDD and DM. We argue that data miners should be familiar with statistical themes and models and statisticians should be aware of the capabilities and limitation of data mining and the ways in which data mining differs from traditional statistics.

Statistics is the traditional field that deals with the quantification, collection, analysis, interpretation, and drawing conclusions from data. Data mining is an interdisciplinary field that draws on computer sciences (data base, artificial intelligence, machine learning, graphical and visualization models), statistics and engineering (pattern recognition, neural networks). DM involves the analysis of large existing data bases in order to discover patterns and relationships in the data, and other findings (unexpected, surprising, and useful). Typically, it differs from traditional statistics on two issues: the size of the data set and the fact that the data were initially collected for purpose other than the that of the DM analysis. Thus, *experimental design*, a very important topic in traditional statistics, is usually irrelevant to DM. On the other hand asymptotic analysis, sometimes criticized in statistics as being irrelevant, becomes very relevant in DM.

While in traditional statistics a data set of 100 to 10^4 entries is considered large, in DM even 10^4 may be considered a small set set fit to be used as an example, rather than a problem encountered in practice. Problem sizes of 10^7 to 10^{10} are more typical. It is important to emphasize, though, that data set sizes are not all created equal. One needs to distinguish between the number of cases (observations) in a large data set (n), and the number of features (variables) available for each case (m). In a large data set,n ,m or both can be large, and it does matter which, a point on which we will elaborate in the continuation. Moreover these definitions may change when the same data set is being used for two different purposes. A nice demonstration of such an instance can be found in the 2001 KDD competition, where in one task the number of cases was the number of purchasing customers, the click information being a subset of the features, and in the other task the clicks were the cases.

Our aim in this chapter is to indicate certain focal areas where statistical thinking and practice have much to offer to DM. Some of them are well known, whereas others are not. We will cover some of them in depth, and touch upon others only marginally. We will address the following issues which are highly relevant to DM:

- Size
- Curse of Dimensionality
- Assessing uncertainty
- Automated analysis
- Algorithms for data analysis in Statistics
- Visualization
- Scalability
- Sampling
- Modelling relationships
- Model selection

We briefly discuss these issues in the next section and then devote special sections to three of them. In section 3 we explain and present how the most basic of statistical methodologies, namely regression analysis, has developed over the years to create a very flexible tool to model relationships, in the form of Generalized Linear Models (GLMs). In section 4 we discuss the False Discovery Rate (FDR) as a scalable approach to hypothesis testing. In section 5 we discuss how FDR ideas contribute to flexible model selection in GLM. We conclude the chapter by asking whether the concepts and methods of KDD and DM differ from those of traditional statistical, and how statistics and DM should act together.

2. Statistical Issues in DM

2.1 Size of the Data and Statistical Theory

Traditional statistics emphasizes the mathematical formulation and validation of a methodology, and views simulations and empirical or practical evidence as a less form of validation. The emphasis on rigor has required proof that a proposed method will work prior to its use. In contrast, computer science and machine learning use experimental validation methods. In many cases mathematical analysis of the performance of a statistical algorithm is not feasible in a specific setting,

but becomes so when analyzed asymptotically. At the same time, when size becomes extremely large, studying performance by simulations is also not feasible. It is therefore in settings typical of DM problems that asymptotic analysis becomes both feasible and appropriate. Interestingly, in classical asymptotic analysis the number of cases n tends to infinity. In more contemporary literature there is a shift of emphasis to asymptotic analysis where the number of variables m tends to infinity. It is a shift that has occurred because of the interest of statisticians and applied mathematicians in wavelet analysis (see Chapter ...), where the number of parameters (wavelet coefficients) equals the number of cases, and has proved highly successful in areas such as the analysis of gene expression data from microarrays.

2.2 The curse of dimensionality and approaches to address it

The curse of dimensionality is a well documented and often cited fundamental problem. Not only do algorithms face more difficulties as the the data increases in dimension, but the structure of the data itself changes. Take, for example, data uniformly distributed in a high-dimensional ball. It turns out that (in some precise way, see Meilijson, 1991) most of the data points are very close to the surface of the ball. This phenomenon becomes very evident when looking for the k-Nearest Neighbors of a point in high-dimensional space. The points are so far away from each other that the radius of the neighborhood becomes extremely large.

The main remedy offered for the curse of dimensionality is to use only part of the available variables per case, or to combine variables in the data set in a way that will summarize the relevant information with fewer variables. This dimension reduction is the essence of what goes on in the data warehousing stage of the DM process, along with the cleansing of the data. It is an important and time-consuming stage of the DM operations, accounting for 80-90% of the time devoted to the analysis.

The dimension reduction comprises two types of activities: the first is quantifying and summarizing information into a number of variables, and the second is further reducing the variables thus constructed into a workable number of combined variables. Consider, for instance, a phone company that has at its disposal the entire history of calls made by a customer. How should this history be reflected in just a few variables? Should it be by monthly summaries of the number of calls per month for each of the last 12 months, such as their means (or medians), their

maximal number, and a certain percentile? Maybe we should use the mean, standard deviation and the number of calls below two standard deviations from the mean? Or maybe we should use none of these but rather variables capturing the monetary values of the activity? If we take this last approach, should we work with the cost itself or will it be more useful to transfer the cost data to the log scale? Statistical theory and practice have much to offer in this respect, both in measurement theory, and in data analysis practices and tools. The variables thus constructed now have to be further reduced into a workable number of combined variables. This stage may still involve judgmental combination of previously defined variables, such as cost per number of customers using a phone lines, but more often will require more automatic methods such as principal components or independent components analysis (for a further discussion of principle component analysis see Roberts and Everson, 2001).

We cannot conclude the discussion on this topic without noting that occasionally we also start getting the *blessing of dimensionality*, a term coined by David Donoho (Donoho, 2000) to describe the phenomenon of the high dimension helping rather than hurting that we often encounter as we proceed up the scale in working with very high dimensional data. For example, for large m if the data we study is pure noise, the i-th largest observation is very close to its expectations under the model for the noise! Another case in point is microarray analysis, where the many non-relevant genes analyzed give ample information about the distribution of the noise, making it easier to identify real discoveries. We shall see a third case below.

2.3 Assessing uncertainty

Assessing the uncertainty surrounding knowledge derived from data is recognized as a the central theme in statistics. The concern about the uncertainty is down-weighted in KDD, often because of the myth that all relevant data is available in DM. Thus, standard errors of averages, for example, will be ridiculously low, as will prediction errors. On the other hand experienced users of DM tools are aware of the variability and uncertainty involved. They simply tend to rely on seemingly "non-statistical" technologies such as the use of a training sample and a test sample. Interestingly the latter is a methodology widely used in statistics, with origins going back to the 1950s. The use of such validation methods, in the form of cross-validation for smaller data sets, has been a common practice in exploratory data analysis when dealing with medium size data sets.

Some of the insights gained over the years in statistics regarding the use of these tools have not yet found their way into DM. Take, for example, data on food store baskets, available for the last four years, where the goal is to develop a prediction model. A typical analysis will involve taking a random training sample from the data, then testing the model on the training sample, with the results guiding us as to the choice of the most appropriate model. However, the model will be used next year, not last year. The main uncertainty surrounding its conclusions may not stem from the person to person variability captured by the differences between the values in the training sample, but rather follow from the year to year variability. If this is the case, we have all the data, but only four observations. The choice of the data for validation and training samples should reflect the higher sources of variability in the data, by each time setting the data of one year aside to serve as the source for the test sample (for an illustrated yet profound discussion of these issues in exploratory data analysis see Mosteller and Tukey, 1977, Ch. 7,8).

2.4 Automated analysis

The inherent dangers of the necessity to rely on automatic strategies for analyzing the data, another main theme in DM, have been demonstrated again and again. There are many examples where trivial non-relevant variables, such as case number, turned out to be the best predictors in automated analysis. Similarly, variables displaying a major role in predicting a variable of interest in the past, may turn out to be useless because they reflect some strong phenomenon not expected to occur in the future (see for example the conclusions using the onion metaphor from the 2002 KDD competition). In spite of these warnings, it is clear that large parts of the analysis should be automated, especially at the warehousing stage of the DM.

This may raise new dangers. It is well known in statistics that having even a small proportion of outliers in the data can seriously distort its numerical summary. Such unreasonable values, deviating from the main structure of the data, can usually be identified by a careful human data analyst, and excluded from the analysis. But once we have to warehouse information about millions of customers, summarizing the information about each customer by a few numbers has to be automated and the analysis should rather deal automatically with the possible impact of a few outliers.

Statistical theory and methodology supply the framework and the tools for this endeavor. A numerical summary of the data that is not

unboundedly influenced by a negligible proportion of the data is called a resistant summary. According to this definition the average is not resistant, for even one straying data value can have an unbounded effect on it. In contrast, the median is resistant. A resistant summary that retains its good properties under less than ideal situations is called a robust summary, the α-trimmed mean (rather than the median) being an example of such. The concepts of robustness and resistance, and the development of robust statistical tools for summarizing location, scale, and relationships, were developed during the 1970's and the 1980's, and resulting theory is quite mature (see, for instance, Ronchetti (Ronchetti et al., 1986; Dell'Aquila and Ronchetti, 2004), even though robustness remains an active area of contemporary research in statistics. Robust summaries, rather than merely averages, standard deviations, and simple regression coefficients, are indispensable in DM. Here too, some adaptation of the computations to size may be needed, but efforts in this direction are being made in the statistical literature.

2.5 Algorithms for data analysis in statistics

Computing has always been a fundamental to statistic, and it remained so even in times when mathematical rigorousity was most highly valued quality of a data analytic tool. Some of the important computational tools for data analysis, rooted in classical statistics, can be found in the following list: efficient estimation by maximum likelihood, least squares and least absolute deviation estimation, and the EM algorithm; analysis of variance (ANOVA, MANOVA, ANCOVA), and the analysis of repeated measurements; nonparametric statistics; log-linear analysis of categorial data; linear regression analysis, generalized additive and linear models, logistic regression, survival analysis, and discriminant analysis; frequency domain (spectrum) and time domain (ARIMA) methods for the analysis of time series; multivariate analysis tools such as factor analysis, principal component and later independent component analyses, and cluster analysis; density estimation, smoothing and denoising, and classification and regression trees (decision trees); Bayesian networks and the Monte Carlo Markov Chain (MCMC) algorithm for Bayesian inference.

For an overview of most of these topics, with an eye to the DM community see Hastie, Tibshirani and Friedman, 2001. Some of the algorithms used in DM which were not included in classical statistic, are considered by some statisticians to be part of statistics (Friedman, 1998). For example, rule induction (AQ, CN2, Recon, etc.), associate rules, neural net-

works, genetic algorithms and self-organization maps may be attributed to classical statistics.

2.6 Visualization

Visualization of the data and its structure, as well as visualization of the conclusions drawn from the data, are another central theme in DM. Visualization of quantitative data as a major activity flourished in the statistics of the 19th century, faded out of favor through most of the 20th century, and began to regain importance in the early 1980s. This importance in reflected in the development of the Journal of Computational and Graphical Statistics of the American Statistical Association. Both the theory of visualizing quantitative data and the practice have dramatically changed in recent years. Spinning data to gain a 3-dimensional understanding of pointclouds, or the use of projection pursuit are just two examples of visualization technologies that emerged from statistics.

It is therefore quite frustrating to see how much KDD software deviates from known principles of good visualization practices. Thus, for instance the fundamental principle that the retinal variable in a graphical display (length of line, or the position of a point on a scale) should be proportional to the quantitative variable it represents is often violated by introducing a dramatic perspective. Add colors to the display and the result is even harder to understand.

Much can be gained in DM by mining the knowledge about visualization available in statistics, though the visualization tools of statistics are usually not calibrated for the size of the data sets commonly dealt within DM. Take for example the extremely effective Boxplots display, used for the visual comparisons of batches of data. A well-known rule determines two fences for each batch, and points outside the fences are individually displayed. There is a traditional default value in most statistical software, even though the rule was developed with batches of very small size in mind (in DM terms). In order to adapt the visualization technique for routine use in DM, some other rule which will probably be adaptive to the size of the batch should be developed. As this small example demonstrates, visualization is an area where joint work may prove to be extremely fruitful.

2.7 Scalability

In machine learning and data mining *scalability* relates to the ability of an algorithm to scale up with size, an essential condition being that the storage requirement and running time should not become infeasible as the size of the problem increases. Even simple problems like multivariate

histograms become a serious task, and may benefit from complex algorithms that scale up with size. Designing scalable algorithms for more complex tasks, such as decision tree modeling, optimization algorithms, and the mining of association rules, has been the most active research area in DM. Altogether, scalability is clearly a fundamental problem in DM mostly viewed with regard to its algorithmic aspects. We want to highlight the duality of the problem by suggesting that concepts should be scalable as well. In this respect, consider the general belief that hypothesis testing is a statistical concept that has nothing to offer in DM. The usual argument is that data sets are so large that every hypothesis tested will turn out to be statistically significant - even if differences or relationships are minuscule. Using association rules as an example, one may wonder whether an observed lift for a given rule is "really different from 1", but then find that at the traditional level of significance used (the mythological 0.05) an extremely large number of rules are indeed significant. Such findings brought David Hand (Hand, 1998) to ask "what should replace hypothesis testing?" in DM. We shall discuss two such important scalable concepts in the continuation: the testing of multiple hypotheses using the False Discovery Rate and the penalty concept in model selection.

2.8 Sampling

Sampling is the ultimate scalable statistical tool: if the number of cases n is very large the conclusions drawn from the sample depend only on the size of the sample and not on the size of the data set. It is often used to get a first impression of the data, visualize its main features, and reach decisions as to the strategy of analysis. In spite of its scalability and usefulness sampling has been attacked in the KDD community for its inability to find very rare yet extremely interesting pieces of knowledge.

Sampling is a very well developed area of statistics (see for example Cochran, 1977), but is usually used in DM at the very basic level. Stratified sampling, where the probability of picking a case changes from one stratum to another, is hardly ever used. But the questions are relevant even in the simplest settings: should we sample from the few positive responses at the same rate that we sample from the negative ones? When studying faulty loans, should we sample larger loans at a higher rate? A thorough investigations of such questions, phrased in the realm of particular DM applications may prove to be very beneficial.

Even greater benefits might be realized when more advanced sampling models, especially those related to super populations, are utilized in DM. The idea here is that the population of customers we view each

year, and from which we sample, can itself be viewed as a sample of the same super population. Hence next year's customers will again be a population sampled from the super population. We leave this issue wide open.

3. Modeling Relationships using Regression Models

Demonstrating that statistics, like data mining, is concerned with turning data into information and knowledge, even though the terminology may differ, in this section we present a major statistical approach being used in data mining, namely regression analysis. In the late 1990s, statistical methodologies such as regression analysis were not included in commercial data mining packages. Nowadays, most commercial data mining software includes many statistical tools and in particular regression analysis. Although regression analysis may seem simple and anachronistic, it is a very powerful tool in DM with large data sets, especially in the form of the generalized linear models (GLMs). We emphasize the assumptions of the models being used and how the underlying approach differs from that of machine learning. The reader is referred to McCullagh and Nelder, 1991 and Chapters ... for more detailed information on the specific statistical methods.

3.1 Linear Regression Analysis

Regression analysis is the process of determining how a variable y is related to one, or more, other variables $x_1, \ldots, x_k$. The y is usually called the *dependent* variable and the x_i's are called the *independent* or *explanatory* variables. In a linear regression model we assume that

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} + \varepsilon_i \quad i = 1, \ldots, M \tag{1.1}$$

and that the ε_i's are independent and are identically distributed as $\mathcal{N}(0, \sigma^2)$ and M is the number of data points. The expected value of y_i is given by

$$E(y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \tag{1.2}$$

To estimate the coefficients of the linear regression model we use the least square estimation which gives results equivalent to the estimators obtained by the maximum likelihood method. Note that for the linear regression model there is an explicit formula of the β's. We can write

(1.1) in matrix form by $Y = X \cdot \boldsymbol{\beta}^t + \boldsymbol{\varepsilon}^t$ where $\boldsymbol{\beta^t}$ is the transpose of the vector $[\beta_0, \beta_1, \ldots, \beta_k]$, $\boldsymbol{\varepsilon^t}$ is the transpose of the vector $\boldsymbol{\varepsilon}= [\varepsilon_1, \ldots, \varepsilon_M]$ and the matrix X is given by

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \cdots & x_{Mk} \end{pmatrix} \tag{1.3}$$

The estimates of the β's are given (in matrix form) by $\hat{\boldsymbol{\beta}}=(X^tX)^{-1}X^tY$. Note that in linear regression analysis we assume that for a given $x_1, \ldots, x_k$ y_i is distributed as $\mathcal{N}(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \sigma^2)$. There is a large class of general regression models where the relationship between the y_is and the vector x is not assumed to be linear, that can be converted to a linear model.

Machine learning approach compared to regression analysis aims to select a function $f \in \mathcal{F}$ from a given set of functions $\mathcal{F}$, that best approximates or fits the given data. Machine learning assumes that the given data $(\mathbf{x}_i, y_i)$, $(i = 1, \ldots, M)$ is obtained by a data generator, producing the data according to an unknown distribution $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$. Given a loss function $\Psi(y - f(\mathbf{x}))$, the quality of an approximation produced by the machine learning is measured by the expected loss, the expectation being below the unknown distribution $p(\mathbf{x}, y)$. The subject of statistical machine learning is the following optimization problem:

$$\min_{f \in \mathcal{F}} \int \Psi(y - f(\mathbf{x}))dp(\mathbf{x}, y) \tag{1.4}$$

when the density function $p(\mathbf{x}, y)$ is unknown but a random independent sample of $(\mathbf{x}_i, y_i)$ is given. The problem of minimizing (1.4) on the basis of the data is the subject of statistical machine learning. If $\mathcal{F}$ is the set of all linear function of $\mathbf{x}$ and $\Psi(y - f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ then if $p(y|\mathbf{x})$ is normally distributed then the minimization of (1.4) is equivalent to linear regression analysis.

3.2 Generalized Linear Models

Although in many cases the set of linear function is good enough to model the relationship between the stochastic response y as a function of $\mathbf{x}$ it may not always suffice to represent the relationship. The generalized linear model increases the family of functions $\mathcal{F}$ that may represent the relationship between the response y and $\mathbf{x}$. The tradeoff is between

having a simple model and a more complex model representing the relationship between y and $\mathbf{x}$. In the general linear model the distribution of y given $\mathbf{x}$ does not have to be normal, but can be any of the distributions in the exponential family (see McCullagh and Nelder, 1991). Instead of the expected value of $y|\mathbf{x}$ being a linear function, we have

$$g(E(y_i)) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji} \tag{1.5}$$

where $g(\cdot)$ is a monotone differentiable function.

In the *generalized additive models*, $g(E(y_i))$ need not to be a linear function of $\mathbf{x}$ but has the form:

$$g(E(y_i)) = \beta_0 + \sum_{j=1}^{k} \sigma_j(x_{ji}) \tag{1.6}$$

where $\sigma(\cdot)$'s are smooth functions. Note that neural networks are a special case of the generalized additive linear models. For example the function that a multilayer feedforward neural network with one hidden layer computes is (see Chapter ... for detailed information):

$$y_i = f(\mathbf{x}) = \sum_{l=1}^{m} \beta_j \cdot \sigma \left(\sum_{j=1}^{k} \mathbf{w}_{jl} \mathbf{x}_{ji} - \theta_j \right) \tag{1.7}$$

where m is the number of processing-units in the hidden layer. The family of functions that can be computed depends on the number of neurons in the hidden layer and the activation function σ. Note that a standard multilayer feedforward network with a smooth activation function σ can approximate any continuous function on a compact set to any degree of accuracy if and only if the network's activation function σ is not a polynomial (Leshno et al., 1993).

There are methods for fitting generalized additive models. However, unlike linear models for which there exits a framework of statistical inference, for machine learning algorithms as well as generalized additive methods, no such framework have yet been developed. For example, using a statistical inference framework in linear regression one can test the hypothesis that all or part of the coefficients are zero.

The total sum of squares (SST) is equal to the sum of squares due to regression (SSR) plus the residual sum of square (RSS_k), i.e.

$$\underbrace{\sum_{i=1}^{M}(y_i-\overline{y})^2}_{SST}=\underbrace{\sum_{i=1}^{M}(\hat{y}_i-\overline{y})^2}_{SSR}+\underbrace{\sum_{i=1}^{M}(y_i-\hat{y}_i)^2}_{RSS_k} \tag{1.8}$$

The percentage of variance explained by the regression is a very popular method to measure the goodness-of-fit of the model. More specifically R^2 and the adjusted R^2 defined below are used to measure the goodness of fit.

$$R^2=\frac{\sum_{i=1}^{M}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{M}(y_i-\overline{y})^2}=1-\frac{RSS_k}{SST} \tag{1.9}$$

$$\text{Adjusted-}R^2=1-(1-R^2)\frac{M-1}{M-k-1} \tag{1.10}$$

We next turn to a special case of the general additive model that is very popular and powerful tool in cases where the responses are binary values.

3.3 Logistic regression

In logistic regression the y_is are binary variables and thus not normally distributed. The distribution of y_i given $\mathbf{x}$ is assumed to follow a binomial distribution such that:

$$log\left(\frac{p(y_i=1|\mathbf{x})}{1-p(y_i=1|\mathbf{x})}\right)=\beta_0+\sum_{j=1}^{k}\beta_j x_{ji} \tag{1.11}$$

If we denote $\pi(\mathbf{x})=p(y=1|\mathbf{x})$ and the real valued function $g(t)=\frac{t}{1-t}$ then $g(\pi(\mathbf{x}))$ is a linear function of $\mathbf{x}$. Note that we can write $y=\pi(\mathbf{x})+\varepsilon$ such that if $y=1$ then $\varepsilon=1-\pi(\mathbf{x})$ with probability $\pi(\mathbf{x})$, and if $y=0$ then $\varepsilon=-\pi(\mathbf{x})$ with probability $1-\pi(\mathbf{x})$. Thus, $\pi(\mathbf{x})=E(y|\mathbf{x})$ and

$$\pi(\mathbf{x})=\frac{e^{\beta_0+\sum_{j=1}^{k}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{k}\beta_j x_j}} \tag{1.12}$$

Of the several methods to estimates the β's, the method of maximum likelihood is one most commonly used in the logistic regression routine of the major software packages.

In linear regression, interest focuses on the size of R^2 or adjusted-R^2. The guiding principle in logistic regression is similar: the comparison of observed to predicted values is based on the log likelihood function. To

compare two models - a full model and a reduced model, one uses the following likelihood ratio:

$$D = -2\ln\left(\frac{\text{likelihod of the reduced model}}{\text{likelihod of the full model}}\right) \tag{1.13}$$

The statistic D in equation (1.13), is called the deviance (McCullagh and Nelder, 1991). Logistic regression is a very powerful tool for classification problems in discriminant analysis and is applied in many medical and clinical research studies.

3.4 Survival analysis

Survival analysis addresses the question of how long it takes for a particular event to happen. In many medical applications the most important response variable often involves time; the event is some hazard or death and thus we analyze the patient's survival time. In business application the event may be a failure of a machine or market entry of a competitor. There are two main characteristics of survival analysis that make it different from regression analysis. The first is that the presence of *censored observation*, where the event (e.g. death) has not necessarily occurred by the end of the study. Censored observation may also occur when patients are lost to follow-up for one reason or another. If the output is censored, we do not have the value of the output, but we do have some information about it. The second, is that the distribution of survival times is often skewed or far from normality. These features require special methods of analysis of survival data, two functions describing the distribution of survival times being of central importance: the *hazard* function and the *survival* function. Using T to represent survival time, the *survival* function denoted by $S(t)$, is defined as the probability of survival time to be greater than t, i.e. $S(t) = \Pr(T > t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function of the output. The hazard function, $h(t)$, is defined as the probability density of the output at time t conditional upon survival to time t, that is $h(t) = f(t)/S(t)$, where $f(t)$ is the probability density of the output. Is is also known as the instantaneous failure rate and presents the probability that an event will happen in a small time interval Δt, given that the individual has survived up to the beginning of this interval, i.e. $h(t) = lim_{\Delta t \downarrow 0} \frac{\Pr(t \leq T < t+\Delta t | t \leq T)}{\Delta t} = f(t)/S(t)$. The hazard function may remain constant, increase, decrease or take some more complex shape. Most modeling of survival data is done using a propor-

tional hazard model. A proportional-hazard model, which assumes that the hazard function is of the form

$$h(t) = \alpha(t) \exp\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right) \tag{1.14}$$

$\alpha(t)$ is a hazard function on its own, called the baseline hazard function, corresponding to that for the average value of all the covariates $x_1, \ldots, x_n$. This is called a proportional-hazard model, because the hazard function for two different patients have a constant ratio. The interpretation of the β's in this model is that the effect is multiplicative.

There are several approaches to survival data analysis. The simplest it to assume that the baseline hazard function is constant which is equivalent to assuming exponential distribution. Another simple approach would be to assume that the baseline hazard function is of the two-parameter family of function, like the Weibull distribution. In these cases the standard methods such as maximum likelihood can be used. In other cases one may restrict $\alpha(t)$ for example by assuming it to be monotonic. In business application, the baseline hazard function can be determined by experimentation, but in medical situations it is not practical to carry out an experiment to determine the shape of the baseline hazard function. The Cox proportional hazards model (Cox, 1972), introduced to overcome this problem, has become the most commonly used procedure for modelling the relationship of covariates to a survival outcome and it is used in almost all medical analyses of survival data. Estimation of the β's is based on the partial likelihood function introduced by Cox (Cox, 1972; Therneau and Grambsch, 2000).

There are many other important statistical themes that are highly relevant to DM, among them: statistical classification methods, spline and wavelets, decision trees and others (see Chapters ... for more detailed information on these issues). In the next section we elaborate on the False Discovery Rate (FDR) method (Benjamini and Hochberg, 1995), a most salient feature of DM.

4. False Discovery Rate (FDR) Control in Hypotheses testing

As noted before there is a feeling that the testing of a hypothesis is irrelevant in DM. However the problem of separating a real phenomenon from its background noise is just as fundamental a concern in DM as in statistics. Take for example an association rule, with an observed lift which is bigger than 1, as desired. Is it also significantly bigger than 1 in the statistical sense, that is beyond what is expected to happen as a

result of noise? The answer to this question is given by the testing of the hypothesis that the lift is 1. However, in DM a hypothesis is rarely tested alone, as the above point demonstrates. The tested hypothesis is always a member of a larger family of similar hypotheses, all association rules of at least a given support and confidence being tested simultaneously. Thus, the testing of hypotheses in DM always invokes the "Multiple Comparisons Problem" so often discussed in statistics which is interesting in itself as the first DM problem in the statistics of 50 years ago did just that: when a feature of interest (a variable) is measured on 10 subgroups (treatments), and the mean values are compared to some reference value (such as 0), the problem is a small one, but take these same means and search among all pairwise comparisons between the treatments to find a significant difference, and the number of comparisons increases to 10*(10-1)/2=45 - which is in general quadratic in the number of treatments. It becomes clear that if we allow an .05 probability of deciding that a difference exists in a single comparison even if it really does not, thereby making a false discovery (or a type I error in statistical terms), we can expect to find on the average 2.25 such errors in our pool of discoveries. No wonder this DM activity is sometimes described in statistics as "post hoc analysis" - a nice definition for DM with a traditional flavor.

The attitude that has been taken during 45 years of statistical research is that in such problems the probability of making even one false discover should be controlled, that is controlling the Family Wise Error rate (FWE) as it is called. The simplest way to address the multiple comparisons problem, and offer FWE control at some desired level α, is to use the Bonferroni procedure: conduct each of the m tests at level α/m. In problems where m becomes very large the penalty to the researcher from the extra caution becomes heavy, in the sense that the probability of making any discovery becomes very small, and so it is not uncommon to avoid the need to adjust for multiplicity.

The False Discovery Rate (FDR), namely the expectation of the proportion of false discoveries (rejected true null hypotheses) among the discoveries (the rejected hypotheses), was developed by Benjamini and Hochberg, 1995 to bridge these two extremes. When the null hypothesis is true for all hypotheses - the FDR and FWE criteria are equivalent. However, when there are some hypotheses for which the null hypotheses are false, an FDR controlling procedure may yield many more discoveries at the expense of having a small proportion of false discoveries.

Formally, let H_{0i}, $i = 1, \ldots m$ be the tested null hypotheses. For $i = 1, \ldots m_0$ the null hypotheses are true, and for the remaining $m_1 = m - m_0$ hypotheses they are not. Thus, any discovery about a hypothesis

from the first set is a false discovery, while a discover about a hypothesis from the second set is a true discovery. Let V denote the number of false discoveries and R the total number of discoveries. Let the proportion of false discoveries be

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0 \end{cases},$$

and define $FDR = E(Q)$.

Benjamini and Hochberg advocated that the FDR should be controlled at some desirable level q, while maximizing the number of discoveries made. They offered the linear step-up procedure as a simple and general procedure that controls the FDR. The linear step-up procedure makes use of the m p-values, $\mathbf{P} = (P_1, \ldots P_m)$ so in a sense it is very general, as it compares the ordered values $P_{(1)} \leq \ldots \leq P_{(m)}$ to the set of constants linearly interpolated between q and q/m.:

Definition 4.1 *The Linear step-up Procedure: Let* $k = \max\{i : P_{(i)} \leq iq/m\}$*, and reject the* k *hypotheses associated with* $P_{(1)}, \ldots P_{(k)}$*. If no such a* k *exists reject none.*

The procedure was first suggested by Eklund (Seeger, 1968) and forgotten, then independently suggested by Simes (Simes, 1986). At both points in time it went out of favor because it does not control the FWE. Benjamini and Hochberg, 1995, showed that the procedure does control the FDR, raising the interest in this procedure. Hence it is now referred to as the Benjamini and Hochberg procedure (BH procedure), or (unfortunately) the FDR procedure (e.g. in SAS). Here, we use the descriptive term, i.e. the linear step-up procedure (for a detailed historical review see Benjamini and Hochberg, 2000).

For the purpose of practical interpretation and flexibility in use, the results of the linear step-up procedure can also be reported in terms of the FDR adjusted p-values. Formally, the FDR adjusted p-value of $\mathrm{H}_{(i)}$ is $p^{LSU}_{(i)} = \min\{\frac{mp_{(j)}}{j} \mid j \geq i\}$. Thus the linear step-up procedure at level q is equivalent to rejecting all hypotheses whose FDR adjusted p-value is $\leq q$.

It should also be noted that the dual linear step-down procedure, which uses the same constants but starts with the smallest p-value and stops at the last $\{P_{(i)} \leq iq/m\}$, also controls the FDR (Sarkar, 2002). Even though it is obviously less powerful, it is sometimes easier to calculate in very large problems.

The linear step-up procedure is quite striking in its ability to control the FDR at precisely $q \cdot m_0/m$, regardless of the distributions of the test statistics corresponding to false null hypotheses (when the distributions under the simple null hypotheses are independent and continuous).

Benjamini and Yekutieli (Benjamini and Yekutieli, 2001) studied the procedure under dependency. For some type of positive dependency they showed that the above remains an upper bound. Even under the most general dependence structure, where the FDR is controlled merely at level $q(1 + 1/2 + 1/3 + \ldots + 1/m)$, it is again conservative by the same factor m_0/m (Benjamini and Yekutieli, 2001).

Knowledge of m_0 can therefore be very useful in this setting to improve upon the performance of the FDR controlling procedure. Were this information to be given to us by an "oracle", the linear step-up procedure with $q' = q \cdot m/m_0$ would control the FDR at precisely the desired level q in the independent and continuous case. It would then be more powerful in rejecting many of the hypotheses for which the alternative holds. In some precise asymptotic sense, Genovese and Wasserman (Genovese and Wasserman, 2002a) showed it to be the best possible procedure.

Schweder and Spjotvoll (Schweder and Spjotvoll, 1982) were the first to try and estimate this factor, albeit informally. Hochberg and Benjamini (Hochberg and Benjamini, 1990) formalized the approach. Benjamini and Hochberg (Benjamini and Hochberg, 2000) incorporated it into the linear step-up procedure, and other adaptive FDR controlling procedures make use of other estimators (see Efron et al., 2001; Storey, 2002, and Storey, Taylor and Siegmund, 2004). Benjamini, Krieger and Yekutieli, 2001, offer a very simple and intuitive two-stage procedure based on the idea that the value of m_0 can be estimated from the results of the linear step-up procedure itself, and prove it controls the FDR at level q.

Definition 4.2 *Two-Stage Linear Step-Up Procedure (TST):*

1. *Use the linear step-up procedure at level* $q' = \frac{q}{1+q}$. *Let* r_1 *be the number of rejected hypotheses. If* $r_1 = 0$ *reject no hypotheses and stop; if* $r_1 = m$ *reject all* m *hypotheses and stop; or otherwise*
2. *Let* $\hat{m}_0 = (m - r_1)$.
3. *Use the linear step-up procedure with* $q^* = q' \cdot m/\hat{m}_0$

Recent papers have illuminated the FDR from many different points of view: asymptotic, Bayesian, empirical Bayes, as the limit of empirical processes, and in the context of penalized model selection (Efron et al., 2001; Storey, 2002; Genovese and Wasserman, 2002a; Abramovich et al., 2001). Some of the studies have emphasized variants of the FDR, such as its conditional value given some discovery is made (the positive FDR in Storey, 2002), or the distribution of the proportion of false discoveries itself (the FDR in Genovese and Wasserman, 2002a; Genovese and Wasserman, 2002b).

Studies on FDR methodologies have become a very active area of research in statistics, many of them making use of the large dimension of the problems faced, and in that respect relying on the blessing of dimensionality. FDR methodologies have not yet found their way into the practice and theory of DM, though it is our opinion that they have a lot to offer there, as the following example on variable selection shows

Example: Zytkov and Zembowicz, 1997; Zembowicz and Zytkov, 1996, developed the 49er software to mine association rules using chi-square tests of significance for the independence assumption, i.e. by testing whether the lift is significantly > 1. Finding that too many of the m potential rules are usually significant, they used $1/m$ as a threshold for significance, comparing each p-value to the threshold, and choosing only the rules that pass the threshold. Note that this is a Bonferroni-like treatment of the multiplicity problem, controlling the FWE at $\alpha = 1$. Still, they further suggest increasing the threshold if a few hypotheses are rejected. In particular they note that the performance of the threshold is especially good if the largest p-value of the selected k rules is smaller than k times the original $1/m$ threshold. This is exactly the BH procedure used at level $q = 1$ and they arrived at it by merely checking the actual performance on a specific problem. In spite of this remarkable success, theory further tells us that it is important to use $q < 1/2$, and not 1, to always get good performance. The preferable values for q are, as far as we know, between 0.05 and 0.2. Such values for q further allow us to conclude that only approximately q of the discovered association rules are not real ones. With $q = 1$ such a statement is meaningless.

5. Model (Variables or Features) Selection using FDR Penalization in GLM

Most of commonly used variable selection procedures in linear models choose the appropriate subset by minimizing a model selection criterion of the form: $RSS_k + \sigma^2 k\lambda$, where RSS_k is the residual sum of squares for a model with k parameters as defined in the previous section, and λ is the penalization parameter. For the generalized linear models discussed above twice the logarithm of the likelihood of the model takes on the role of RSS_k, but for simplicity of exposition we shall continue with the simple linear model. This penalized sum of squares might ideally be minimized over all k and all subsets of variables of size k, but practically in larger problems it is usually minimized either by forward selection or backward elimination, adding or dropping one variable at a time. The different selection criteria can be identified by the value of λ they

use. Most traditional model selection criteria make use of a fixed λ and can also be described as fixed level testing. The Akaike Information Criterion (AIC) and the C_p criterion of Mallows both make use of $\lambda = 2$, and are equivalent to testing at level 0.16 whether the coefficient of each newly included variable in the model is different than 0. Usual backward and forward algorithms use similar testing at the .05 level, which is approximately equivalent to using $\lambda = 4$.

Note that when the selection of the model is conducted over a large number of potential variables m, the implications of the above approach can be disastrous. Take for example $m = 500$ variables, not an unlikely situation in DM. Even if there is no connection whatsoever between the predicted variable and the potential set of predicting variables, you should expect to get 65 variables into the selected model - an unacceptable situation.

More recently model selection approaches have been examined in the statistical literature in settings where the number of variables is large, even tending to infinity. Such studies, usually held under an assumption of orthogonality of the variables, have brought new insight into the choice of λ. Donoho and Jonhstone (Donoho and Johnstone, 1995) suggested using λ, where $\lambda = 2log(m)$, whose square root is called the "universal threshold" in wavelet analysis. Note that the larger the pool over which the model is searched, the larger is the penalty per variable included. This threshold can also be viewed as a multiple testing Bonferroni procedure at the level α_m, with $.2 \leq \alpha_m \leq .4$ for $10 \leq m \leq 10000$. More recent studies have emphasized that the penalty should also depend on the size of the already selected model k, $\lambda = \lambda_{k,m}$, increasing in m and decreasing in k. They include Abramovich and Benjamini, 1996; Birge and Massart, 2001; Abramovich, Bailey and Sapatinas , 2000; Tibshirani and Knight, 1999; George and Foster, 2000, and Foster and Stine, 2004. As full review is beyond our scope, we shall focus on the suggestion that is directly related to FDR testing.

In the context of wavelet analysis Abramovich and Benjamini, 1996 suggested to using FDR testing, thereby introducing a threshold that increases in m and decreases with k. Abramovich, Bailey and Sapatinas , 2000, were able to prove in an asymptotic setup, where m tends to infinity and the model is sparse, that using FDR testing is asymptotically minimax in a very wide sense. Their argument hinges on expressing the FDR testing as a penalized RSS as follows:

$$RSS_k + \sigma^2 \sum_{i=1}^{i=k} z^2_{\frac{i}{m} \cdot \frac{q}{2}}, \tag{1.15}$$

where z_α is the $1-\alpha$ percentile of a standard normal distribution. This is equivalent to using $\lambda_{k,m} = \frac{1}{k}\sum_{i=1}^{i=k} z^2_{\frac{i}{m}\cdot\frac{q}{2}}$ in the general form of penalty. When the models considered are sparse, the penalty is approximately $2\sigma^2 log(\frac{m}{k}\cdot\frac{2}{q})$. The FDR level controlled is q, which should be kept at a level strictly less than 1/2.

In a followup study Gavrilov, 2003, investigated the properties of such penalty functions using simulations, in setups where the number of variables is large but finite, and where the potential variables are correlated rather than orthogonal. The results show the dramatic failure of all traditional "fixed penalty per-parameter" approaches. She found the FDR-penalized selection procedure to have the best performance in terms of minimax behavior over a large number of situations likely to arise in practice, when the number of potential variables was more than 32 (and a close second in smaller cases). Interestingly they recommend using $q = .05$, which turned out to be well calibrated value for q for problems with up to 200 variables (the largest investigated).

Example: Foster and Stine, 2004, developed these ideas for the case when the predicted variable is 0-1, demonstrating their usefulness in DM, in developing a prediction model for loan default. They started with approximately 200 potential variables for the model, but then added all pairwise interactions to reach a set of some 50,000 potential variables. Their article discusses in detail some of the issues reviewed above, and has a very nice and useful discussion of important computational aspects of the application of the ideas in real a large DM problem.

6. Concluding Remarks

KDD and DM are a vaguely defined field in the sense that the definition largely depends on the background and views of the definer. Fayyad defined DM as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Some definitions of DM emphasize the connection of DM to databases containing ample of data. Another definitions of KDD and DM is the following definition: "Nontrivial extraction of implicit, previously unknown and potentially useful information from data, or the search for relationships and global patterns that exist in databases". Although mathematics, like computing is a tool for statistics, statistics has developed over a long time as a subdiscipline of mathematics. Statisticians have developed mathematical theories to support their methods and a mathematical formulation based on probability theory to quantify the uncertainty. Traditional statistics emphasizes a mathematical formulation and validation of its methodology rather than empirical or practical validation.

The emphasis on rigor has required a proof that a proposed method will work prior to the use of the method. In contrast, computer science and machine learning use experimental validation methods. Statistics has developed into a closed discipline, with its own scientific jargon and academic objectives that favor analytic proofs rather than practical methods for learning from data. We need to distinguish between the theoretical mathematical background of statistics and its use as a tool in many experimental scientific research studies. We believe that computing methodology and many of the other related issues in DM should be incorporated into traditional statistics. An effort has to be made to correct the negative connotations that have long surrounded data mining in the statistics literature (Chatfield, 1995) and the statistical community will have to recognize that empirical validation does constitute a form of validation (Friedman, 1998).

Although the terminology used in DM and statistics may differ, in many cases the concepts are the same. For example, in neural networks we use terms like "learning", "weights" and "knowledge" while in statistics we use "estimation", "parameters" and "value of parameters", respectively. Not all statistical themes are relevant to DM. For example, as DM analyzes existing databases, experimental design is not relevant to DM. However, many of them, including those covered in this chapter, are highly relevant to DM and any data miner should be familiar with them.

In summary, there is a need to increase the interaction and collaboration between data miners and statistics. This can be done by overcoming the terminology barriers, working on problems stemming from large databases. A question that has often been raised among statisticians is whether DM is not merely part of statistics. The point of this chapter was to show how each can benefit from the other, making the inquiry from data a more successful endeavor, rather than dwelling on theoretical issues of dubious value.

References

Abramovich F. and Benjamini Y., (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics & Data Analysis*, 22:351–361.

Abramovich F., Bailey T .C. and Sapatinas T., (2000). Wavelet analysis and its statistical applications. *Journal of the Royal Statistical Society Series D-The Statistician*, 49:1–29.

Abramovich F., Benjamini Y., Donoho D. and Johnstone I., (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000-19, Department of Statistics, Stanford University.

Benjamini Y. and Hochberg Y., (1995). Controlling the false discover rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Benjamini Y. and Hochberg Y., (2000). On the adaptive control of the false discovery fate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25:60–83.

Benjamini Y., Krieger A.M. and Yekutieli D., (2001). Two staged linear step up for controlling procedure. Technical report, Department of Statistics and O.R., Tel Aviv University.

Benjamini Y. and Yekutieli D., (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Berthold M. and Hand D., (1999). *Intelligent Data Analysis: An Introduction*. Springer.

Birge L. and Massart P., (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3:203–268.

Chatfield C., (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society A*, 158:419–466.

Cochran W.G., (1977). *Sampling Techniques*. Wiley.

Cox D.R., (1972). Regressio models and life-tables. *Journal of the Royal Statistical Society B*, 34:187–220.

Dell'Aquila R. and Ronchetti E.M., (2004). *Introduction to Robust Statistics with Economic and Financial Applications*. Wiley.

Donoho D.L. and Johnstone I.M., (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90:1200–1224.

Donoho D., (2000). American math. society: Math challenges of the 21st century: *High-dimensional data analysis: The curses and blessings of dimensionality*.

Efron B., Tibshirani R.J., Storey J.D. and Tusher V., (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160.

Friedman J.H., (1998). *Data Mining and Statistics: What's the connections?*, Proc. 29th Symposium on the Interface (D. Scott, editor).

Foster D.P. and Stine R.A., (2004). Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99:303–313.

Gavrilov Y., (2003). Using the falls discovery rate criteria for model selection in linear regression. M.Sc. Thesis, Department of Statistics, Tel Aviv University.

Genovese C. and Wasserman L., (2002a). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B*, 64:499–517.

Genovese C. and Wasserman L., (2002b). A stochastic process approach to false discovery rates. Technical Report 762, Department of Statistics, Carnegie Mellon University.

George E.I. and Foster D.P., (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–748.

Hand D., (1998). Data mining: Statistics and more? *The American Statistician*, 52:112–118.

Hand D., Mannila H. and Smyth P., (2001). *Principles of Data Mining*. MIT Press.

Han J. and Kamber M., (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publisher.

Hastie T., Tibshirani R. and Friedman J., (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hochberg Y. and Benjamini Y., (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9:811–818.

Leshno M., Lin V.Y., Pinkus A. and Schocken S., (1993). Multilayer feedforward networks with a non polynomial activation function can approximate any function. *Neural Networks*, 6:861–867.

McCullagh P. and Nelder J.A., (1991). *Generalized Linear Model*. Chapman & Hall.

Meilijson I., (1991). The expected value of some functions of the convex hull of a random set of points sampled in r^d. *Isr. J. of Math.*, 72:341–352.

Mosteller F. and Tukey J.W., (1977). *Data Analysis and Regression : A Second Course in Statistics*. Wiley.

Roberts S. and Everson R. (editors), (2001). *Independent Component Analysis : Principles and Practice*. Cambridge University Press.

Ronchetti E.M., Hampel F.R., Rousseeuw P.J. and Stahel W.A., (1986). *Robust Statistics : The Approach Based on Influence Functions*. Wiley.

Sarkar S.K., (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, 30:239–257.

Schweder T. and Spjotvoll E., (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502.

Seeger P., (1968). A note on a method for the analysis of significances en mass. *Technometrics*, 10:586–593.

Simes R.J., (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754.

Storey J.D., (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B*, 64:479–498.

Storey J.D., Taylor J.E. and Siegmund D., (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society Series B*, 66:187–205.

Therneau T.M. and Grambsch P.M., (2000). *Modeling Survival Data, Extending the Cox Model*. Springer.

Tibshirani R. and Knight K., (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society Series B*, 61:Part 3 529–546.

Zembowicz R. and Zytkov J.M., (1996). From contingency tables to various froms of knowledge in databases. In U.M. Fayyad, R. Uthurusamy, G. Piatetsky-Shapiro and P. Smyth (editors) *Advances in Knowledge Discovery and Data Mining* (pp. 329-349). MIT Press.

Zytkov J.M. and Zembowicz R., (1997). Contingency tables as the foundation for concepts, concept hierarchies and rules: The 49er system approach. *Fundamenta Informaticae*, 30:383–399.