# Toward a Statistical Foundation for Data Mining

Norman Matloff
Department of Computer Science
University of California at Davis
Davis, CA 95616

T.Y. Lin
Department of Computer Science
San Jose State University
San Jose, CA 95192 USA

## Abstract

*KDD is an inherently statistical activity, and there has been considerable literature which draws upon statistical science. However, the usage has typically been vague and informal at best, and at worst of a seriously misleading nature. The present paper seeks to take a first step in remedying this problem by pairing precise mathematical descriptions of the concepts in KDD with practical interpretations and implications for specific KDD issues.*

## 1 Introduction

A number of papers by statisticians have noted the inherently statistical nature of KDD, such as Pregibon [1], Friedman [2] and Rocke [3]. These authors have pointed out various statistical tools which should prove useful in KDD. Yet these papers have not taken a mathematically rigorous approach in their presentation.

Similarly, much of the research in KDD has been of a purely empirical nature, without a supporting theoretical framework. As the old saying goes, "The proof of the pudding is in the eating," so empirical evaluation is a must for any method. Yet it is important also to *understand* why a method works well or not, and this calls for a mathematical treatment at some level.

On the other hand, there are many papers which make the opposite error: They are quite rigorous but fail to connect to real-world issues in a practical, intuitive manner. Moreover, their theoretical nature renders them largely inaccessible to KDD practitioners.

Clearly, this gap—a gaping chasm, really—is not healthy for the field. We believe the gap is resulting in both misunderstandings and missed opportunities. Our aim here, then, will be to bridge this gap—or more accurately, to advocate that researchers in the field themselves bridge the gap, along the lines we propose in this paper:

We call upon empirical researchers to couch the problems and solutions they discuss in a mathematically precise manner, and at least undertake a small-scale mathematical analysis. At the same time, we call upon theoretical researchers to present their work in a more intuitive manner with a more solid connection to actual KDD practice.

In both cases, we will propose a simple framework, consisting of some simple mathematical constructs motivated by intuitive notions tied to the actual practice of KDD. It is important to note that the latter, i.e. the intuitive "philosophical" issues, will play an integral role here. The mathematical constructs are not sufficient by themselves. They will be simple, and in fact will be at least vaguely familiar to some readers, but our emphasis here will be in their interpretation and usage.

In Section 2, we will set up the mathematical framework and their intuitive interpretation in a practical context. Then in the following sections we present several examples of published work in KDD which we believe would have been enhanced by the precise yet intuitive approach we have outlined above. We will argue, for example, that contrary to having a constraining effect on empirical research, a precise yet intuitive formulation of the issues in a research project can actually ehance the researcher's ability to do innovative, "out of the box" thinking.

## 2 Some Infrastructure

As is common in theoretical treatments, we will phrase the issues in terms of a statistical prediction problem. But we depart from tradition by engaging in an explicit discussion of the practical interpretation of what we mean by "statistical."

### 2.1 Notation

Denote our attribute set by $X^{(1)}, ..., X^{(d)}$. It is assumed that our database constitutes a *statistical sample* of n observations on these attributes; the $i^{th}$ observation on the $j^{th}$ attribute from this sample is denoted by $X_i^{(j)}$, i = 1,...,n, j = 1,...,d.

To make things concrete—again, this is one of our principle aims—let's consider the well-known KDD "basket" example. Each row in the database corresponds to some individual consumer. Some of the attributes may be characteristics of a consumer, say age, income or gender (say 1 for male, 0 for female), while others will record whether the consumer bought certain items (1 for yes, 0 for no).

The vector $(X_i^{(1)}, ..., X_i^{(d)})$, representing the values in the $i^{th}$ observation of all our attributes will be denoted by $X_i$. In relational database terms, this vector is the $i^{th}$ row in our relation.

### 2.2 Sampling from Populations, Real or Conceptual

In considering our database to be a "statistical sample," we mean that it is a sample from some "population." This interpretation is, in our view, key. The population may be tangible, as in the "basket" example, where we are sampling from the population of all customers of this business.

Or, the population may be more conceptual in nature. A database consisting of students in a new major in a university could be considered as a sample from the conceptual population of all students at this university, who might be in this major. If for example we imagine the university overall enrollment had been 20 percent larger this year, with no change in demographic or other makeup of the enrollment, then some of the increased overall enrollment would have been students choosing this major.

Here is an example of a "population" which is even more conceptual in nature. Consider the subject of quadratic equations, studied in school algebra classes:

$$ax^2 + bx + c = 0 \tag{1}$$

The students learn that this equation has a real root if and only the *discriminant* $b^2 - 4ac$ is nonnegative. Suppose one did not know this rule, and tried to find it using KDD.

This sounds like an inherently non-statistical problem. Yet one could convert it to a statistical problem in the following way. One could sample randomly from $a/b/c$ space, and for each triplet from this space, determine somehow (say by graphing the quadratic polynomial) whether a real root exists. One could then apply various *statistical regression models* (see below), trying to predict the 0-1 variable $w$ from $a$, $b$ and $c$. In this manner, we might possibly stumble on the discriminant rule. discriminant rule

### 2.3 Relation to Probability Distributions

It is important to relate the abstract mathematical variables to the population being studied. When we speak of the distribution of $X^{(j)}$, what we really mean is the distribution of that attribute in the population. Say $X^{(1)}$ is age of the customer. When we say, for instance, that $P(X^{(1)} > 32) = 0.22$, we mean that 22 percent of all customers in this population are older than 32.

A similar point holds for expected value. Some KDD practitioners with an engineering background might be accustomed to interpreting $E(X^{(1)})$ in terms of the physics metaphor of center of gravity. Yet for statistical applications such as KDD, the relevant interpretation of this quantity is as the mean age of all customers in this population.

## 3 Prediction

The essence of the statistical nature of KDD is prediction. For notational convenience, in the remainder of this paper, let us suppose that we are using $X^{(1)}, ..., X^{(d-1)}$ to predict $X^{(d)}$, and rename the latter variable $Y$.

Our focus here will be on predicting *dichotomous*, i.e. 0/1-valued, variables Y here. (We do not make this restriction on the variables $X^{(j)}$.)

### 3.1 Statement of the Problem

Suppose for the moment that we know the population distributions of the attributes, and we wish to minimize the overall probability of misclassification.[1] Suppose that we observe $X^{(j)}$ to have the value $v_j$, j = 1,...,d-1. Then we would guess $Y$ to be either 0 or 1, according to whether

$$q(v) = q(v_1, ..., v_{d-1}) = P(Y = 1|X^{(1)} = v_1, ..., X^{(d-1)} = v_{d-1}) \quad (2)$$

is less than 0.5 or greater than 0.5, respectively, where $v = q(v_1, ..., v_{d-1})$).

### 3.2 Classification Vs. Regression

Some authors, e.g. Han [], consider the case of dichotomous $Y$ to be a conceptually separate case from that of continuous $Y$, and refer to it as *classification* instead of prediction, but mathematically it is the same problem, in the following sense.

Classically, the problem of predicting a general variable $Y$ from a vector of attributes $X = (X^{(1)}, ..., X^{(d-1)})$ is posed as finding a function h() that minimizes

$$E[(Y - h(X))^2] \quad (3)$$

One can easily show that the solution is the *regression function*, h(t) defined by

$$h(t) = E(Y|X = t) \quad (4)$$

Now, if $Y$ is dichotomous, i.e. $Y$ takes on the values 0 and 1, then

$$E(Y|X = t) = 1 \cdot P(Y = 1|X = t) + 0 \cdot P(Y = 0|X = t) = q(t) \quad (5)$$

In other words, the general formulation of the prediction problem yields the function q() anyway.

This is not just a semantic issue. A vast literature exists on the general regression problem, with much relevant material,[2] and it would be a loss not to draw upon it. Note by the way that the computation in the case of the logistic classification model described below is done via (nonlinear) regression algorithms.[3]

### 3.3 The Function q() Must Be Estimated from Sample Data

This is complicated by the fact that we do not the population distributions of the attributes, as assumed in the previous paragraph. We thus do not know the function $q()$ above, and need to estimate it from the observations in our database.

The estimated function, $\hat{q}(v)$, is obtained either by parametric or nonparametric means. A common parametric approach, for instance, uses the logistic regression model [], which postulates that $q(v)$ has the form

$$q(v_1, ..., v_{d-1}) = \frac{1}{1 + exp[-(\beta_0 + \beta_1 v_1 + ... + \beta_{d-1} v_{d-1})]}$$

The parameters $\beta_j$ are estimated from our sample data $X_i^{(j)}$, yielding the estimated parameters $\hat{\beta}_j$ and the estimated q(v):

[1] The latter would not be the case if we assigned different costs to different types of errors. It may be more costly to false guess $Y$ to be 1 than to false guess it to be 0, for example.

[2] A large separate literature on the classification problem has also been developed, but much of it draws upon the material on regression.

[3] See for example the **lrm** procedure in the R statistical package []. By the way, some of these points are also noted (albeit rather abstractly) in Friedman [].

$$\hat{q}(v) = \frac{1}{1 + exp[-(\hat{\beta}_0 + \hat{\beta}_1 v_1 + ... + \hat{\beta}_{d-1} v_{d-1})]}$$

A nonparametric method often used in KDD for estimating q(v) is CART [].

Note that we will make a classification error if $q(v)$ and $\hat{q}(v)$ are on opposite sides of the value 0.5.

### 3.4 "Noise Mining"

There is a lot of talk about "noise mining," "overfitting" and the like in the KDD literature, but again this is rarely precisely defined.

In some cases, the literature does point out that the "average" discrepancy between $\hat{q}(v)$ and $q(v)$ can be shown to consist of two components—a *bias* component,

$$E[(E\hat{q}(v) - q(v))^2] \tag{6}$$

and a *variance* component,

$$E[(\hat{q}(v) - E\hat{q}(v))^2] \tag{7}$$

Note that $v$ is fixed here, not random. Instead, the randomness involves the fact that these expected values are averages over all possible samples from the given population.[4] This interpretation is very important when one is assessing various competing types of prediction methodology, and especially important in understanding the bias/variance problem (Section 3.4).

A large bias is due to using too simple a model in the parametric case, or to using too large a granularity in the nonparametric case (e.g. leaf nodes too large in CART). In both cases, one common source of the problem is that we are using too few predictor attributes.

However, any efforts to reduce the bias will increase the variance, i.e. increase the amount of "noise." This is due to having an insufficient sample size n for the given model. In CART, for example, a given rectangle might contain very few observations, thus rendering $\hat{q}()$ inaccurate within the rectangle. The same rectangle, applied to a larger sample from the same population, might work fine.

Clearly, there is a tradeoff between bias and variance for fixed n. As finer models are fitted, the bias is reduced but the variance increases. If too much attention is paid to minimizing bias rather than variance, the decision rules found from the analysis may be spurious, hence the term *noise mining*.

## 4 Worked-Out Example

Recall that our theme here has been that empirical research work in KDD should include a mathematically precise statement of the problem, and present mathematical treatment of at least a small but illustrative model of the effects being studied. In that light, we now present such a model of the "noise fitting" problem.

Continue to assume the setting described at the beginning of Section 3, but with the additional specialization that all the predictor attributes $X^{(j)}$, j = 1,...,d-1 are dichotomous.

Suppose that $X^{(j)}$, j = 1,...,d-1 all "coin tosses," i.e. have probability 0.5 of taking on the value 1 and are statistically independent. Suppose in addition that $P(Y = 1|X^{(1)} = v_1)$ is equal to 0.6 for $v_1 = 1$ and equal to 0.4 for $v_1 = 0$.

Under these circumstances, $X^{(j)}$, j = 2,...,d-1 have no predictive power for Y at all, and

$$q(v_1, v_2, ..., v_{d-1}) = P(Y = 1|X^{(1)} = v_1) \tag{8}$$

independent of $v_2, ..., v_{d-1}$.

But we would not know this, since we would not have the population data. We would have only *sample* estimates of q(v) to work with, $\hat{q}(v)$. The point then is that that estimate will be subject to the bias/variance issues discussed here. We discuss the variance issue first, and focus our attention on the estimation of q(1,1,...,1).

[4] So, our "population" here can be viewed as the n-fold cartesian product of the original population, with the expected value being an average over all points in that meta-population.

One decision we would need to make is which of the attributes $X^{(j)}$ to use as predictors. Let us compare the effects of using just $X^{(1)}$ alone to predict Y, versus using $X^{(1)}, X^{(2)}, ..., X^{(d-1)}$ for that prediction. In the former situation, note again that we would be modeling q(v) to be a function which does not depend on $X^{(2)}, ..., X^{(d-1)}$ (see Equation (8)). Again, this modeling assumption would be correct, but we would not know this.

Suppose we are not using a parametric model, and instead are simply using straight sample proportions to estimate q(). Then if we use only $X^{(1)}$ as our predictor, as discussed above, our estimate of q(1,1,...,1) would be the proportion of records in our database for which Y = 1, among those for which $X^{(1)} = 1$, i.e.

$$\hat{q}(1,1,...,1) = \frac{\sum_i X_i^{(1)} X_i^{(d)}}{\sum_i X_i^{(1)}} = \frac{T_1}{U_1} \tag{9}$$

The question at hand is, "What is the probability that $\hat{q}()$ will make the right decision for us in this situation, which is to guess that Y = 1?"[5] Well, this is

$$P(\hat{q}(1,1,...,1) > 0.5) = P(T_1 > 0.5U_1) \tag{10}$$

To evaluation this probability, note that just as $T_1$ and $U_1$, being binomially distributed,[6] have approximate normal distributions, their bivariate distribution approximates that of a bivariate normal.[7] The means and variances of $T_1$ and $U_1$ are then np, nq, np(1-p) and nq(1-q), where $p = P(X^{(1)} = X^{(d)} = 1) = 0.3$ and $q = P(X^{(1)} = 1) = 0.5$. The covariance is

$$\begin{aligned} & Cov(T_1, U_1) \\ &= n[E(X^{(1)}X^{(d)}X^{(1)}) - E(X^{(1)}X^{(d)}) \cdot EX^{(1)}] \\ &= np(1-q) \end{aligned} \tag{11}$$

Any linear combination of $T_1$ and $U_1$, say $aT_1 + bU_1$, then has an approximate normal distribution with mean n(ap+bq), and variance

$$a^2 Var(T_1) + b^2 Var(U_1) + 2abCov(T_1, U_1) \tag{12}$$

In our case here, a = 1 and b = -0.5. After doing the calculations we find that $E(T_1 - 0.5U_1) = -0.05n$ and $Var(T_1 - 0.5U_1) = 0.1225n$,
and thus

$$P(T_1 > 0.5U_1) \approx P(Z > -0.64\sqrt{n}) \tag{13}$$

where Z is a standard N(0,1) variate.

So, Equation (9) is the probability that we make the right decision if we predict Y from only $X^{(1)}$. Let's see how that probability changes if we predict Y from $X^{(1)}, ..., X^{(d-1)}$.

In this setting, Equation (8) reverts to (2), and (9) becomes

$$\hat{q}(1,1,...,1) = \frac{\sum_i X_i^{(1)} X_i^{(2)} ... X_i^{(d-1)} X_i^{(d)}}{\sum_i X_i^{(1)} X_i^{(2)} ... X_i^{(d-1)}} = \frac{T_{d-1}}{U_{d-1}} \tag{14}$$

Equation (13) then becomes (after a bit of algebraic approximation)

$$P(T_{d-1} > 0.5U_{d-1}) \approx P(Z > -0.5^{d/2}\sqrt{n}) \tag{15}$$

Compare Equations (13) and (15), focusing on the roles of d and n. They are both of the form P(Z ¿ c) for a negative c, and the algebraically smaller (i.e. more negative) c is, the better. So, for fixed n, the larger d is, the worse is our predictive ability for Y.

[5] This is the right decision in the sense that it is the best guess given $X^{(1)}$. It doesn't necessarily mean that that guess will be correct.

[6] The variable $W = X^{(1)}X^{(d)}$ is 0-1 valued, so the sum $T_1$ is binomial.

[7] This stems from the fact the vector form of the Central Limit Theorem.

Now, remember the context: We devised a model here in which $X_i^{(2)}...X_i^{(d-1)}$ had no predictive ability at all for Y in the population distribution, though the analyst would not know this. *In other words, not only will the analyst not gain predictive ability by using these attributes, he/she would actually lose predictive power by using them, i.e. we "overfit."*

So, this is the variance side of the bias/variance tradeoff. The number of records in our sample which have $X^{(1)} = 1, X^{(2)} = 1, ..., X^{(d-1)} = 1$ will be very small for large d (similar to having a small leaf node in CART), leading to a high variance for $\hat{q}(1, 1, ..., 1)$.

Equation (15) also shows the role of n in the overfitting issue: For fixed d, as n increases the harmful effect of overfitting will diminish.

Now, what about the bias side of the bias/variance tradeoff? Suppose we are considering using $X^{(1)}, X^{(2)}..., X^{(k)}$ as our predictor attributes. Due to the nature of the model here, the bias in using any k in the range $1 \leq k < d - 1$ is 0. So, if we use k greater than 1, we are incurring the problem of increasing variance without reducing bias.

On the other hand, using k = 0 would produce a bias, since $X^{(1)}$ does have some predictive value for Y: If k were taken to be 0, then the population value of q(1,1,...,1) would reduce to the unconditional probability P(Y = 1) = 0.5, rather than the true value 0.6.

Again, our point in devising this model here is to illustrate our theme that even empirical KDD research should anchor its presentation with (a) a precise mathematical statement of the problem being studied, and (b) a simple mathematical model which explicitly illustrates the issues.

The word *explicitly* in (b) should be emphasized. Equation (15) explicitly shows the roles of d and n. One sees that for a fixed value of n, use of a larger d increases the variance, possibly more than the bias is reduced. As d increases, at some point our predictive ability based on sample data will begin to diminish, i.e. we will overfit. One also sees, though, that for a larger value of n, that crossover point will occur for a larger d, i.e. we can use more attributes as our predictors.