Clark Glymour,
David Madigan, Daryl Pregibon,
and Padhraic Smyth

Statistics may have little to offer the search architectures in a data mining search, but a great deal to offer in evaluating hypotheses in the search, in evaluating the results of the search, and in applying the results.

Statistical Inference and Data Mining

DATA MINING AIMS TO DISCOVER SOMETHING NEW FROM THE FACTS RECORDED in a database. For many reasons—encoding errors, measurement errors, unrecorded causes of recorded features—the information in a database is almost always noisy; therefore, inference from databases invites applications of the theory of probability. From a statistical point of view, databases are usually uncontrolled convenience samples; therefore data mining poses a collection of interesting, difficult—sometimes impossible—inference problems, raising many issues, some well studied and others unexplored or at least unsettled.

Data mining almost always involves a search architecture requiring evaluation of hypotheses at the stages of the search, evaluation of the search output, and appropriate use of the results. Statistics has little to offer in understanding search architectures but a great deal to offer in evaluation of hypotheses in the course of a search, in evaluating the results of a search, and in understanding the appropriate uses of the results.

TERRY WIDENER

Here we describe some of the central statistical ideas relevant to data mining, along with a number of recent techniques that may sometimes be applied. Our topics include features of probability distributions, estimation, hypothesis testing, model scoring, Gibb's sampling, rational decision making, causal inference, prediction, and model averaging. For a rigorous survey of statistics, the mathematically inclined reader should see [7]. Due to space limitations, we must also ignore a number of interesting topics, including time series analysis and meta-analysis.

Probability Distributions

The statistical literature contains mathematical characterizations of a wealth of probability distributions, as well as properties of random variables—functions defined on the "events" to which a probability measure assigns values. Important relations among probability distributions include marginalization (summing over a subset of values) and conditionalization (forming a conditional probability measure from a probability measure on a sample space and some event of positive probability. Essential relations among random variables include independence, conditional independence, and various measures of dependence—of which the most famous is the correlation coefficient. The statistical literature also characterizes families of distributions by properties useful in identifying any particular member of the family from data, or by closure properties useful in model construction or inference (e.g., conjugate families closed under conditionalization and the multinormal family closed under linear combination). Knowledge of the properties of distribution families can be invaluable in analyzing data and making appropriate inferences.

Heuristic procedures, which abound in machine learning (and in statistics), have no guarantee of ever converging on the right answer.

Inference involves the following features:

- Estimation
- Consistency
- Uncertainty
- Assumptions
- Robustness
- Model averaging

Many procedures of inference can be thought of as estimators, or functions from data to some object to be estimated, whether the object is the values of a parameter, intervals of values, structures, decision trees, or something else. Where the data are a sample from a larger (actual or potential) collection described by a probability distribution for any given sample size, the array of values of an estimator over samples of that size has a probability distribution. Statistics investigates such distributions of estimates to identify features of an estimator related to the information, reliability, and uncertainty it provides.

An important feature of an estimator is consistency; in the limit, as the sample size increases without bound, estimates should almost certainly converge to the correct value of whatever is being estimated. Heuristic procedures, which abound in machine learning (and in statistics), have no guarantee of ever converging on the right answer. An equally important feature is the uncertainty of an estimate made from a finite sample. That uncertainty can be thought of as the probability distribution of estimates made from hypothetical samples of the same size obtained in the same way. Statistical theory provides measures of uncertainty (e.g., standard errors) and methods of calculating them for various families of

estimators. A variety of resampling and simulation techniques have also been developed for assessing uncertainties of estimates [1]. Other things (e.g., consistency) being equal, estimators that minimize uncertainty are preferred.

The importance of uncertainty assessments can be illustrated in many ways. For example, in recent research aimed at predicting the mortality of hospitalized pneumonia patients, a large medical database was divided into a training set and a test set. (Search procedures used the training set to form a model, and the test set helped assess the predictions of the model.) A neural net using a large number of variables outperformed several other methods. However, the neural net's performance turned out to be an accident of the particular train/test division. When a random selection of other train/test divisions (with the same proportions) were made and the neural net and competing methods trained and tested according to each, the average neural net performance was comparable to that of logistic regression.

Estimation is almost always made on the basis of a set of assumptions, or model, but for a variety of reasons the assumptions may not be strictly met. If the model is incorrect, estimates based on it are also expected to be incorrect, although that is not always the case. One aim of statistical research is to find ways to weaken the assumptions necessary for good estimation. Robust statistics looks for estimators that work satisfactorily for larger families of distributions; resilient statistics [3] concern estimators—often order statistics—that typically have small errors when assumptions are violated.

A more Bayesian approach to the problem of estimation under assumptions emphasizes that alternative models and their competing assumptions are often plausible. Rather than making an estimate based on a single model, several models can be considered, each with an appropriate probability, and when each of the competing models yields an estimate of the quantity of interest, an estimate can be obtained as the weighted average of the estimates given by the individual models [5]. When the probability weights are well calibrated to the frequencies with which the various models obtain, model averaging is bound to improve estimation on average. Since the models obtained in data mining are usually the result of some automated search procedure, the advantages of model averaging are best obtained if the error frequencies of the search procedure are known—something usually obtainable only through extensive Monte Carlo exploration. Our impression is that the error rates of search procedures proposed and used in the data mining and statistical literatures are rarely estimated in this way. (See [10] and [11] for Monte Carlo test design-for-search procedures.) When the probabilities of various models are entirely subjective, model averaging gives at least coherent estimates.

Hypothesis Testing

Hypothesis testing can be viewed as one-sided estimation in which, for a specific hypothesis and any sample of an appropriate kind, a testing rule either conjectures that the hypothesis is false or makes no conjecture. The testing rule is based on the conditional sampling distribution (conditional on the truth of the hypothesis to be tested) of some statistic or other. The significance level α of a statistical test specifies the probability of erroneously conjecturing that the hypothesis is false (often called rejecting the hypothesis) when the hypothesis is in fact true. Given an appropriate alternative hypothesis, the probability of failing to reject the hypothesis under test can be calculated; that probability is called the power of the test against the alternative. The power of a test is obviously a function of the alternative hypothesis being considered.

Since statistical tests are widely used, some of their important limitations should be noted. Viewed as a one-sided estimation method, hypothesis testing is inconsistent unless the alpha level of the testing rule is decreased appropriately as the sample size increases. Generally, a level α test of one hypothesis and a level α test of another hypothesis do not jointly provide a level α test of the conjunction of the two hypotheses. In special cases, rules (sometimes called contrasts) exist for simultaneously testing several hypotheses [4]. An important corollary for data mining is that the alpha level of a test has nothing directly to do with the probability of error in a search procedure that involves testing a series of hypotheses. If, for example, for each pair of a set of variables, hypotheses of independence are tested at $\alpha = 0.5$, then 0.5 is not the probability of erroneously finding some dependent set of variables when in fact all pairs are independent. That relation would hold (approximately) only when the sample size is much larger than the number of variables considered. Thus, in data mining procedures that use a sequence of hypothesis tests, the alpha level of the tests cannot generally be taken as an estimate of any error probability related to the outcome of the search.

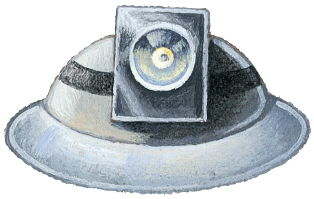
In many, perhaps most, realistic hypothesis spaces, hypothesis testing is comparatively uninformative. If a hypothesis is not rejected by a test rule and a sample, the same test rule and the same sample may very well

also not reject many other hypotheses. And in the absence of knowledge of the entire power function of the test, the testing procedure provides no information about such alternatives. Further, the error probabilities of tests have to do with the truth of hypotheses, not with approximate truth; hypotheses that are excellent approximations may be rejected in large samples. Tests of linear models, for example, typically reject them in very large samples no matter how closely they seem to fit the data.

Model Scoring

THE evidence provided by data should lead us to prefer some models or hypotheses to others and to be indifferent about still other models. A *score* is any rule that maps models and data to numbers whose numerical ordering corresponds to a preference ordering over the space of models, given the data. For such reasons, scoring rules are often an attractive alternative to tests. Indeed, the values of test statistics are sometimes themselves used as scores, especially in the structural-equation literature. Typical rules assign to a model a value determined by the likelihood function associated with the model, the number of parameters, or dimension, of the model, and the data. Popular rules include the Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), and Minimum Description Length. Given a prior probability distribution over models, the posterior probability on the data is itself a scoring function, arguably a privileged one. The BIC approximates posterior probabilities in large samples.

There is a notion of consistency appropriate to scoring rules; in the large sample limit, the true model should almost surely be among those receiving maximal scores. AIC scores are generally not consistent [8]. The probability (p) values assigned to statistics in hypothesis tests of models are scores, but it does not seem to be known whether and under what conditions they form a consistent set of scores. There are also uncertainties associated with scores, since two different samples of the same size from the same distribution can yield not only different numerical values for the same model but even different orderings of models.



When the probabilities of various models are entirely subjective, model averaging gives at least coherent estimates.

For obvious combinatorial reasons, it is often impossible when searching a large model space to calculate scores for all models; however, it is often feasible to describe and calculate scores for a few equivalence classes of models receiving the highest scores.

In some contexts, inferences made using Bayes scores and posteriors can differ a great deal from inferences made with hypothesis tests. (See [5] for examples of models that account for almost all of the variance of an outcome of interest and that have very high posterior or Bayes scores but are overwhelmingly rejected by statistical tests.)

Of the various scoring rules, perhaps the most interesting is the posterior probability, because, unlike many other consistent scores, posterior probability has a central role in the theory of rational choice. Unfortunately, posteriors can be difficult to compute.

Gibbs Sampling

Statistical theory typically gives asymptotic results that can be used to describe posteriors or likelihoods in large samples. Unfortunately, even in very large databases, the number of cases relevant to a particular question can be quite small. For example, in studying the effects of hospitalization on survival of pneumonia patients, mortality comparisons between those treated at home and those treated in a hospital might be wanted. But even in a very large database, the number of pneumonia patients treated at

home and who die of pneumonia complications is very small. And statistical theory typically provides few or no ways to calculate distributions in small samples in which the application of asymptotic formulas can be wildly misleading. Recently, a family of simulation methods—often described as Gibbs sampling after the great American physicist Josiah Willard Gibbs 1839–1903, have been adapted from statistical mechanics, permitting the approximate calculation of many distributions. A review of these procedures is in [9].

Rational Decision Making and Planning

The theory of rational choice assumes the decision maker has a definite set of alternative actions, knowledge of a definite set of possible alternative states of the world, and knowledge of the payoffs or utilities of the outcomes of each possible action in each possible state of the world, as well as knowledge of the probabilities of various possible states of the world. Given all this information, a decision rule specifies which of the alternative actions ought to be taken. A large literature in statistics and economics addresses alternative decision rules—maximizing expected utility, minimizing maximum possible loss, and more. Rational decision making and planning are typically the goals of data mining, but rather than providing techniques or methods for data mining, the theory of rational choice poses norms for the use of information obtained from a database.

The very framework of rational decision making requires probabilities for alternative states of affairs and knowledge of the effects alternative actions will have. To know the outcomes of actions is to know something of cause-and-effect relations. Extracting such causal information is often one of the principal goals of data mining and more generally of statistical inference.

Inference to Causes

Understanding causation is the hidden motivation behind the historical development of statistics. From the beginning of the field, in the work of Bernoulli and Laplace, the absence of causal connection between two variables has been taken to imply their probabilistic independence [12]; the same idea is fundamental in the theory of experimental design. In 1934, Sewell Wright, a biologist, introduced directed graphs to represent causal hypotheses (with vertices as random variables and edges representing direct influences); these graphs have become common representations of causal hypotheses in the social sciences, biology, computer science, and engineering.

In 1982, statisticians Harry Kiiveri and T. P. Speed combined directed graphs with a generalized connection between independence and absence of causal connection in what they called the Markov condition—if Y is not an effect of X, then X and Y are conditionally independent, given the direct causes of X. Kiiveri and Speed showed that much of the linear modeling literature tacitly assumed the Markov condition; the Markov condition is also satisfied by most causal models of categorical data and of virtually all causal models of systems without feedback. Under additional assumptions, conditional independence provides information about causal dependence. The most common—and most thoroughly investigated—additional assumption is that all conditional independencies are due to the Markov condition's being applied to the directed graph describing the actual causal processes generating the data, a requirement with many names (e.g., faithfulness). Directed graphs with associated probability distributions satisfying the Markov condition are called by different names in different literatures (e.g., Bayes nets, belief nets, structural equation models, and path models).

CAUSAL inference from uncontrolled convenience samples is liable to many sources of error. Three of the most important are latent variables (or confounders), sample selection bias, and model equivalence. A latent variable is any unrecorded feature that varies among recorded units and whose variation influences recorded features. The result is an association among recorded features not in fact due to any causal influence of the recorded features themselves. The possibility of latent variables can seldom, if ever, be ignored in data mining. Sample selection bias occurs when the values of any two of the variables under study, say X and Y, themselves influence whether a feature is recorded in a database. That influence produces a statistical association between X and Y (and other variables) that has no causal significance. Datasets with missing values pose sample selection bias problems. Models with quite different graphs may generate the same constraints on probability distributions through the Markov condition and may therefore be indistinguishable without experimental intervention. Any procedure that arbitrarily selects one or a few of the equivalents may badly mislead users when the models are given a causal significance. If model search is viewed as a form of estimation, all of these difficulties are sources of inconsistency.

Standard data mining methods run afoul of these difficulties. The search algorithms in such commercial linear model analysis programs as LISREL select one from an unknown number of statistically indistinguishable models. Regression methods are inconsistent for all of the reasons listed earlier. For example, consider the structure: $Y = aT + ey$; $X1 = bT$

+ cQ + $e1$; $X2 = dQ + e2$, where T and Q are unrecorded. Neither $X1$ nor $X2$ has any influence on Y. For all nonzero values of a, b, c,, d, however, in sufficiently large samples, regression of Y on $X1$, $X2$ yields significant regression coefficients for $X1$ and $X2$. With the causal interpretation often given it, regression says that $X1$ and $X2$ cause of Y. Assuming the Markov and faithfulness conditions, all that can be inferred correctly (in large samples) from data on $X1$, $X2$, and Y is that $X1$ is not a cause of $X2$ or of Y; $X2$ is not a cause of Y; Y is not a cause of $X2$; and there is no common cause of Y and $X2$. Nonregression algorithms implemented in the TETRAD II program [6, 10] give the correct result asymptotically in this case and in all cases in which the Markov and faithfulness conditions hold. The results are also robust against the three problems with causal inference noted in the previous paragraph [11]. However, the statistical decisions made by the algorithms are not really optimal, and the implementations are limited to the multinomial and multinormal families of probability distributions. A review of Bayesian search procedures for causal models is given in [2].

Prediction

Sometimes one is interested in using a sample, or a database, to predict properties of a new sample, where it is assumed the two samples are obtained from the same probability distribution. As with estimation, prediction is interested in accuracy and uncertainty, and is often measured by the variance of the predictor.

Prediction methods for this sort of prediction problem always assume some regularities—constraints—in the probability distribution. In data mining contexts, the constraints are typically either supplied by human experts or automatically inferred from the database. For example, regression assumes a particular functional form for relating variables or, in the case of logistic regression, relating the values of some variables to the probabilities of other variables; but constraints are implicit in any prediction method that uses a database to adjust or estimate the parameters used in prediction. Other forms of constraint may include independence, conditional independence, and higher-order conditions on correlations (e.g., tetrad constraints). On average, a prediction method guaranteeing satisfaction of the constraints realized in the probability distribution is more accurate and has a smaller variance than a prediction method that does not. Finding the appropriate constraints to be satisfied is the most difficult issue in this sort of prediction. As with estimation, prediction can be improved by model averaging, provided the probabilities of the alternative assumptions imposed by the model are available.

Understanding causation is the hidden motivation behind the historical development of statistics.

Another sort of prediction involves interventions that alter the probability distribution—as in predicting the values (or probabilities) of variables under a change in manufacturing procedures or changes in economic or medical treatment policies. Making accurate predictions of this kind requires some knowledge of the relevant causal structure and is generally quite different from prediction without intervention, although the same caveats about uncertainty and model averaging apply. For graphical representations of causal hypotheses according to the Markov condition, general algorithms for predicting the outcomes of interventions from complete or incomplete causal models were developed in [10]. In 1995, some of these procedures were extended and made into a more convenient calculus by Judea Pearl, a computer scientist. A related theory without graphical models was developed in 1974 by Donald Rubin, a statistician, and

others, and in 1986 by James Robins.

Well-known studies by Herbert Needleman, a physician and statistician, of the correlation of lead deposits in children's teeth and the children's IQs resulted, eventually, in removal of tetraethyl lead from gasoline in the U.S. One dataset Needleman examined included more than 200 subjects and measured a large number of covariates. In 1985, Needleman and his colleagues reanalyzed the data using backward stepwise regression of verbal IQ on these variables and obtained six significant regressors, including lead. In 1988, Steven Klepper, an economist, and his collaborators reanalyzed the data assuming that all the variables were measured with error. Klepper's model assumes that each measured number is a linear combination of the true value and an error and that the parameters of interest are not the regression coefficients but the coefficients relating the unmeasured true-value variables to the unmeasured true value of verbal IQ.

These coefficients are in fact indeterminate—or, in econometric terminology, unidentifiable. However, an interval estimate of the coefficients that is strictly positive or negative for each coefficient can be made if the amount of measurement error can be bounded with prior knowledge by an amount that varies from case to case. For example, Klepper found that the bound required to ensure the existence of a strictly negative interval estimate for the lead-to-IQ coefficient was much too strict to be credible; thus he concluded that the case against lead was not nearly as strong as Needleman's analysis suggested.

Allowing the possibility of latent variables, Richard Scheines in 1996 reanalyzed the correlations with the TETRAD II program and concluded that three of the six regressors could have no influence on IQ. The regression included the three extra variables only because the partial regression coefficient is estimated by conditioning on all other regressors—just the right thing to do for linear prediction, but the wrong thing to do for causal inference using the Markov condition (see the example at the end of the earlier section Inference to Causes). Using the Klepper model—but without the three irrelevant variables—and assigning to all of the parameters a normal prior probability with mean zero and a substantial variance, Scheines used Gibbs sampling to compute a posterior probability distribution for the lead-to-IQ parameter. The probability is very high that lead exposure reduces verbal IQ.

Conclusion

The statistical literature has a wealth of technical procedures and results to offer data mining, but it also offers several methodological morals:

- Prove that estimation and search procedures used in data mining are consistent under conditions reasonably thought to apply in applications;
- Use and reveal uncertainty—don't hide it;
- Calibrate the errors of search—for honesty and to take advantages of model averaging;
- Don't confuse conditioning with intervening, that is, don't take the error probabilities of hypothesis tests to be the error probabilities of search procedures.

Otherwise, good luck. You'll need it.

References

1. Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics (SIAM), Number 38, Philadelphia, 1982.
2. Heckerman, D. Bayesian networks for data mining. Data Mining and Knowledge Discovery, submitted.
3. Hoaglin, D., Mosteller, F., and Tukey, J. *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.
4. Miller, R. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 1981.
5. Raftery, A.E. Bayesian model selection in social research. Working Paper 94-12, Center for Studies in Demography and Ecology, Univ. of Washington, Seattle, 1994.
6. Scheines, R., Spirtes, P., Glymour, C., and Meek, C. *TETRAD II: Tools for Causal Modeling. Users Manual*. Erlbaum, Hillsdale, N.J., 1994.
7. Schervish, M. *Theory of Statistics*. Springer-Verlag, New York, 1995.
8. Schwartz, G. Estimating the dimension of a model. *Ann. Stat. 6* (1978), 461–464.
9. Smith, A.F.M., and Roberts, G.O. Bayesian computation via the Gibb's sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc., Series B, 55 (1993), 3–23.
10. Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.
11. Spirtes, P., Meek, C., and Richardson, T. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. P. Besnard and S. Hanks, Eds. Morgan Kaufmann Publishers, San Mateo, Calif., 1995, pp. 499–506.
12. Stigler, S. *The History of Statistics*. Harvard University Press, Cambridge, Mass., 1986.

Additional references for this article can be found at http://www.research.microsoft.com/research/datamine/CACM-DM-refs/.

CLARK GLYMOUR is Alumni University Professor at Carnegie Mellon University and Valtz Family Professor of Philosophy at the University of California, San Diego. He can be reached cg09@andrew.cmu.edu.

DAVID MADIGAN is an associate professor of statistics at the University of Washington in Seattle. He can be reached at madigan@stat.washington.edu.

DARYL PREGIBON is the head of statistics research in AT&T Laboratories. He can be reached at daryl@research.att.com

PADHRAIC SMYTH is an assistant professor of information and computer science at the University of California, Irvine. He can be reached at smyth@ics.uci.edu.