A Statistical Perspective on KDD

John F. Elder IV
Computational and Applied Mathematics Department
& Center for Research on Parallel Computation
Rice University
elder@rice.edu

Daryl Pregibon
Statistics & Data Analysis Research
AT&T Bell Laboratories
daryl@research.att.com

Abstract

The quest to find models usefully characterizing data is a process central to the scientific method and has been carried out on many fronts. Researchers from an expanding number of fields have designed algorithms to discover rules or equations that capture key relationships between variables in a database. Some modern heuristic modeling approaches seem to have gained in popularity partly as a way to "avoid statistics" while still addressing challenging induction tasks. Yet, there are useful distinctives in what may be called a "statistical viewpoint", and we review here some major advances in statistics from recent decades that are applicable to Knowledge Discovery in Databases.[1]

Recent Statistical Contributions

It would be unfortunate if the KDD community dismissed statistical methods on the basis of courses that they took on statistics several to many years ago. The following provides a rough chronology of "recent" significant contributions in statistics that are relevant to the KDD community. The noteworthy fact is that this time period coincides with the significant increases in computing horsepower and memory, powerful and expressive programming languages, and general accessibility to computing that has propelled us into the Information Age. In effect, this started a slow but deliberate shift in the statistical community, whereby important influences and enablers were to come from computing rather than mathematics.

The 1960s

This was the era of *robust* and *resistant* statistical methods. Following ideas of G. E. P. Box and J. W. Tukey, Huber (1964) and Hampel (1974) formalized the notion that the usual estimators of location and regression coefficients were very sensitive to "outliers", "leverage values", and otherwise unreasonably small amounts of contamination. Key concepts are the

- *influence* function of Hampel (essentially the derivative of an estimator with respect to the data),
- *M*-estimators of Huber, so-called because they generalize maximum likelihood estimators (which require a probability distribution) to a closely related class of estimating equations, and
- *diagnostics*, where implicit downweighting of observations afforded by robust estimators is replaced by empirical derivatives that quantify the effects of small changes in the data on important aspects of regression-like models (see for example, Belsley, Kuh, and Welsch, 1980).

The theory supporting these ideas is elegant and important as it unifies many seemingly unrelated concepts (*e.g.* trimmed means and medians) and more so because it reflects the realism that data does not usually obey assumptions as required by (mathematical) theorems. Thus the robustness era freed statisticians of the shackles of narrow models depending on unrealistic assumptions (*e.g.* normality).

The only downside of the era was that too much effort was placed on deriving new estimators that deviated only slightly from each other both qualitatively and quantitatively.[2] What was needed instead, was the leadership and direction in *using* these methods in practice and dealing with the plethora of alternatives available. Partly because of this misguided effort, many of the techniques of the era never made it into commercial software and therefore never made it into the mainstream of methods used by nonstatisticians.

The Early 1970s

The term Exploratory Data Analysis (EDA) characterizes the notion that statistical insights and modeling are driven by data. John Tukey (1977; Mosteller and Tukey, 1977) reinforced these notions in the early

[1] This paper is an excerpt from a chapter to appear in *Advances in Knowledge Discovery and Data Mining* (Fayyad, et al., 1995). Topics not covered here but in the longer paper include a survey of classical and modern modelling and classification algorithms, and discussion of recent advances in statistical computing and graphics.

[2] Basically reflecting R. A. Fisher's insight (*Statistical Methods for Researchers*, 1924) that there is nothing easier than inventing a new statistical estimator.

70's using a battery of ultra-simple methods, *e.g.*, what could be done with pencil and paper. But the deeper message was to dispel the traditional dogma stating that one was not allowed to "look at the data prior to modeling". On the one side the argument was that hypotheses and the like must not be biased by choosing them on the basis of what the data seem to be indicating. On the other side was the belief that pictures and numerical summaries of data are necessary in order to understand how rich a model the data can support.

A key notion in this era characterized statistical modeling as decomposing the data into structure and noise,

$$data = fit + residual \qquad (1)$$

and then examining residuals to identify and move additional structure into the fit. The fitting process would then be repeated and followed by subsequent residual analyses.

The iterative process described above has its roots in the general statistical paradigm of partitioning variability into distinct parts (*e.g.*, explained and unexplained variation; or, in classification, within-group and between-group variation). The EDA notion simply uses the observed scale of the response rather than the somewhat unnatural squared units of "variability". While this might seem like a trivial distinction, the difference is critical since it is only on the observed scale that diagnosis *and* treatment is possible. For example, a component of variance can indicate that nonlinearity is present but cannot prescribe how to accommodate it.

Graphical methods (not to be confused with graphical *models* in Bayes nets) enjoyed a renaissance during this period as statisticians (re-)discovered that nothing outperforms human visual capabilities in pattern recognition. Specifically, statistical tests and models focus on *expected* values, and in many cases, it is the *un*expected that upsets or invalidates a model (*e.g.*, outliers). Tukey argued that (good) graphical methods should allow *un*expected values to present themselves — once highlighted, models can be expanded or changed to account for them.

Another important contribution was to make data *description* respectable once more. Statistics has its roots in earlier times when descriptive statistics reigned and mathematical statistics was only a gleam in the eye. Data description is concerned with simplicity and accuracy, while not being overly formal about quantifying these terms (though an important area of research tries to do just that; *e.g.*, Mallows (1973), Akaike (1973), and Rissanen (1978)). A key notion popularized in this era was that there is seldom a single *right* answer – in nearly all situations there are many answers. Effective data description highlights those that are simple, concise, and reasonably accurate. Simple transformations of a dataset are used to effect such descriptions, the two most common ones being *data reexpression, e.g.,* using $\log(age)$ instead of *age*, and *data splitting, e.g.,* setting aside outliers to simplify the description of the bulk of the data.

The Late 1970s

To an outsider much of the statistical literature would seem fragmented and disjoint. But the fact of the matter is that much is closely related, but that specific details of individual contributions hide the real similarities. In the late 70's, two review papers and one book elegantly captured the essence of numerous prior publications. The first of these, *Generalized Linear Models* (Nelder and Wedderburn, 1974; McCullagh and Nelder, 1989) extended the classical linear model to a much wider class that included probability models other than the normal distribution and structural models that were nonlinear. The theory accomplished this by decomposing the variation in a response variable into systematic and random components, allowing the former to capture covariate effects through a strictly monotone *link* function, $g(\mu) = \sum x_j \beta_j$, and allowing the latter to be a member of the exponential family of distributions, $\mathcal{E}(\mu, \sigma)$. In so doing, these models provided a unifying theory for regression-like models for binary and count data, as well as continuous data from asymmetric distributions. The second major review paper is well known outside of statistics as the *EM* algorithm (Dempster, Laird, and Rubin, 1977). This paper neatly pulled together numerous ways of solving estimation problems with incomplete data. But the beauty of their general treatment was to instill the concept that even if data are complete, it is often useful to treat it as a missing value problem for computational purposes. Finally, the analysis of nominal or discrete data, specifically counts, had several disconnected streams in the literature and inconsistent ways to describe relationships. Bishop, Fienberg, and Holland (1975) pulled this material together into the class of *loglinear* models. The associated theory allowed researchers to draw analogies to models for continuous data (for example, analysis of variance ideas) and further provided computational strategies for estimation and hypothesis testing. It is also noteworthy that this work anticipated current work in so-called *graphical models*, a subset of the class of loglinear models for nominal data.

The Early 1980s

Resampling methods had been around since the late 1950s under the moniker *jackknife*, so-named by Tukey because it was a "trusty general purpose tool" for eliminating low-order bias from an estimator (Schreuder, 1986). The essence of the procedure is to replace the original n observations by n or more (possibly) correlated estimates of the quantity of interest (called *psuedovalues*). These are obtained by systematically leaving out one or more observations and recomputing the estimator. More precisely, if θ is the parameter of in-

terest, the ith psuedovalue is defined by

$$p_i = n\hat{\theta}_{all} - (n-k)\hat{\theta}_{-i} \qquad (2)$$

where the last quantity is the estimator $\hat{\theta}$ based on leaving out the ith subset (of size k). The jackknife estimate of θ is the arithmetic mean of the psuedovalues, $\bar{p} = \sum p_i/n$.

While the jackknife was originally proposed as a bias reduction tool, it was quickly recognized that the ordinary standard deviation of the psuedovalues provides an honest estimate of the error in the estimate. Thus an empirical means of deriving a measure of uncertainty for virtually *any* statistical estimator was available. One interpretation of the procedure is that the construction of psuedovalues is based on repeatedly and systematically sampling *without* replacement from the data at hand. This led Efron (1979) to generalize the concept to repeated sampling *with* replacement, the so-called *bootstrap* (since it allowed one to "pick oneself up by the bootstraps" in constructing a confidence interval or standard error). This seemingly trivial insight opened the veritable flood gates for comprehensive analytic study and understanding of resampling methods. The focus on estimating precision of estimators rather than bias removal coupled with the advance of computing resources, allowed standard errors of highly nonlinear estimators to be routinely considered.

Unfortunately, as with robustness, the bulk of the research effort was directed at theoretical study of resampling ideas in what KDD researchers would regard as uninteresting situations. The most nonlinear procedures, such as those resulting from combining model identification and model estimation, received only cursory effort (*e.g.* Efron and Gong, 1983; Faraway, 1991).

The Late 1980s

One might characterize classical statistical methods as being "globally" linear whereby the explanatory/prediction/classification variables affect the distribution of the response variable via linear combinations. Thus the effect of x_j on y is summarized by a single regression coefficient β_j. Nonlinear relationships could only be modeled by specifically including the appropriate nonlinear terms in the model, *e.g.* x_j^2 or $\log x_j$. Cleveland (1979) helped seed the notion that globally linear procedures could be replaced with locally linear ones by employing scatterplot smoothers in interesting ways. A scatterplot smoother $s(x)$ is a data-dependent curve defined pointwise over the range of x. For example, the *moving average* smoother is defined at each unique x, as the mean $\bar{y}(x) = \sum y_i/k$ of the k (symmetric) nearest neighbors of x. The ordered sequence of these pointwise estimates traces out a "smooth" curve through the scatter of (x, y) points. Originally smoothers were used simply to enhance scatterplots where clutter or changing density of plotted points hindered visual interpretation of trends and nonlinear features. But by interpreting a scatterplot smoother as an estimate of the conditional mean $E(y|x)$, one obtains an adaptive, nonlinear estimate of the effect of x on the distribution of y. Moreover, this nonlinearity could be tamed while simultaneously reducing bias caused by end-effects, by enforcing "local" linearity in the smoothing procedure (as opposed to local constants as provided by moving averages or medians). Thus by moving a *window* across the data and fitting linear regressions within the window, a globally nonlinear fit is obtained, *i.e.* the sequence of predictions at each point x_i, $s_i(x) = a_i + b_i x$, where the coefficients a_i and b_i are determined by the least squares regression of y on x for all points in the window centered on x_i.

This notion has been applied now in many contexts (*e.g.* regression, classification, discrimination) and across many "error" distributions (*e.g.* the generalized additive model of Hastie and Tibshirani, 1985). While this work reduced the emphasis on strict linearity of the explanatory variables in such models, it did not ameliorate the need for having previously identified the relevant variables to begin with.

The Early 1990s

Within the statistics community, Friedman and Tukey (1974) pioneered the notion of allowing a model to adapt even more nonlinearly by letting the data determine the interesting structure present with "projection pursuit" methods. These are less restrictive than related nonlinear methods such as neural networks, supposing a model of the form

$$\mu(y|x) = \sum_{k=1}^{K} g_k\left(\sum_{j=1}^{J} x_j \beta_{jk}\right) \qquad (3)$$

where both the regression coefficients β_{jk} and the *squashing* functions $g_k()$ are unknown.

Important algorithmic developments and theory resulted from these models even though they failed to achieve widespread use within the statistics community. Part of the reason was that these models were regarded as *too* flexible in the sense that arbitrarily complex functions could be provably recovered (with big enough K). The community instead retreated back to additive models that had limited flexibility but afforded much greater interpretability. Indeed, interpretability was the focus of much of the work in this era as alternative formulations of the locally linear model were derived, *e.g.*, penalized likelihood and Bayesian formulations (O'Sullivan *et al.* 1986).

Still, these ambitious methods helped to nudge the community from focusing on model estimation to model selection. For example, with tree-based models (Breiman *et al* 1984) and multiple adaptive regression splines (Friedman 1991), the modeling search is over structure space as well as parameter space. It is not uncommon now for many thousands of candidate

structures to be considered in a modeling run – which forces one to be even more conservative when judging whether improvements are significant, since any measure of model quality optimized by a search is likely to be over-optimistic. When considering a plethora of candidates it usually becomes clear that a wide variety of models, with different structures and even inputs, score nearly as well as the single "best". Current research in Bayesian *model averaging* and *model blending* combines many models to obtain estimates with reduced variance and (almost always) better accuracy on new data (*e.g.*, Wolpert, 1992; Breiman, 1994b). Such techniques seem especially promising when the models being merged are from completely different families (for example, trees, polynomials, kernels, and splines).

Distinctives of Statistical Practice

Researchers from different fields seem to emphasize different qualities in the models they seek. As with workers in Machine Learning and KDD (but unlike most using, say, neural network and polynomial network techniques), Statisticians are usually interested in *interpreting* their models and may sacrifice some performance to be able to extract meaning from the model structure. If the accuracy is acceptable they reason that a model which can be decomposed into revealing parts is often more useful than a "black box" system, especially during early stages of the investigation and design cycle.

Statisticians are also careful to propagate uncertainty (or randomness) in sampled data to estimated models, summarizing the induced randomness by so-called *sampling distributions* of estimators. By judicious assumptions, exact sampling distributions are analytically tractable; more typically asymptotic arguments are invoked. The net result is often the same, the estimated parameters are approximately normally distributed. This distribution characterizes the uncertainty in the estimated parameters, and owing to normality, the uncertainty is succinctly captured in the standard deviation of the sampling distribution, termed the *standard error* of the estimate. Parameters associated with estimates that are small in comparison to their standard errors, (*e.g.*, $t = \hat{\beta}/s.e.(\hat{\beta}) < 2$) are not likely to be part of the "true" underlying process generating the data, and it is often prudent to drop such parameters from the model. [3]

[3]The Bayesian paradigm provides a different though related perspective where one treats the parameter itself as a random variable and merges prior beliefs about the parameter together with observed data. The resulting *posterior* distribution, $p(\theta|data)$, can often itself be approximated by a normal distribution, and thereby a single number summary of parameter uncertainty is available. Of course, recent computational advances and ingenious algorithms (*e.g.* Markov chain monte carlo) obviate the need for analytically derived normal approximations.

An important consideration that statisticians have faced concerns the case where inferences are desired but data is sparse. Consider an example from retail marketing. An SKU (stock keeping unit) is a unique label assigned to a retail product, for example, men's size 12 blue socks. Predictions of SKUs are required at a store level in a large chain of department stores to build up sufficient inventory for promotions and seasonal demand or other "predictable" events. The problem is that detailed historical data on individual SKU sales at each and every store in the chain is not available; for example, it may be that no men's size 12 blue socks sold in the Florida store since last November. The concept of *borrowing strength* allows one to build forecasts at the site-SKU level by exploiting hierarchies in the problem, possibly in more ways than one. By aggregating across stores, sufficient information is available to build a site-independent prediction for each SKU. This prediction can be used to add stability to predictions of SKUs in each of several regions, which can in turn be used to add stability at the site level. Similar types of decompositions could allow us to borrow strength by looking at sales of, say, all blue socks independent of size, then all socks, then men's undergarments, then menswear overall. Such "hierarchical models" have their origins in *empirical Bayes* models, so-called because inferences are not truly Bayesian, as maximum likelihood estimates are used in place of "hyperparameters" (the parameters in prior distributions) at the highest levels of the hierarchy where data is most numerous. This typically results in estimates of the form $\hat{y}_i = \alpha\bar{y}_i + (1-\alpha)\bar{y}$ where $\bar{y}_i$ is the estimate specific to the ith level of the hierarchy and $\bar{y}$ to that of its parent (where data is more abundant). The mixing parameter, α, captures the similarity of the individual estimate to its parent relative to the tightness of the distribution of the $\bar{y}_i$'s.

The tradeoff between model "underfit" (bias) and "overfit" (variance) is a standard one in statistics. If data are plentiful, model overfit can be avoided by reserving representative subsets of the data for testing as the model is constructed. When performance on the test set systematically worsens, model growth is curtailed. With limited data, all the cases can be employed for training but additional information is needed to *regularize* the fit. Some examples of regularization criteria include model complexity (*e.g.*, number of parameters), roughness (*e.g.*, integrated squared slope of its response surface), and parameter *shrinkage*, (*e.g.*, parameters are smoothly shrunken toward zero). The criterion to be minimized is a weighted sum of the training error and regularization criteria: *criterion* = *model-accuracy* $+\alpha\times$ *regularization-penalty.*

The scalar α is most usually chosen using some form of *cross-validation, e.g.,* bootstrap or leave-out-K methods. Most regularization procedures have a Bayesian interpretation whereby the user-defined prior guides the direction and degree of regularization.

Finally, although some of the appeal of non-traditional models and methods undoubtedly stems from their apparent ability to bypass statistical analysis stages many see as cumbersome, it is clear that matching the *assumptions* of a method with the characteristics of a problem is beneficial to its solution. Statistical analysts usually take the useful step of checking those assumptions; chiefly, by examining:

1. residuals (model errors)
2. diagnostics (model sensitivity to perturbations)
3. parameter covariances (redundancy)

Not all violations of assumptions are equally bad. For example, assumptions about stochastic behavior of the data are typically less important than the structural behavior; the former might lead to inefficient estimates or inaccurate standard errors, but the later could result in biased estimates. Within these two broad classes, normality and independence assumptions are typically less important than homogeneity of variance (*e.g.*, $var(y|x) = constant$ for all x). A single *outlier* from the structural model can bias the fit everywhere else. Likewise, *leverage values* are those observations that have undue influence on the fit, for example if deleting the ith observation results in a large change in the estimate of a key parameter. An important distinction is that leverage values need not correspond to large residuals – indeed by virtue of their "leverage", they bias the fit toward themselves resulting in small or negligible residuals. *Colinearity* among the predictor variables confuses the interpretation of the associated parameters, but can also be harmful to prediction; the new data must strictly abide by the interrelationships reflected in the training data or the model will be extrapolating beyond the confines of the training space, rather than interpolating within it.

Reservations to Automatic Modeling

The experienced statistician, perhaps the most capable of guiding the development of automated tools for data analysis, may also be the most acutely aware of all the difficulties that can arise when dealing with real data. This hesitation has bred skepticism of what automated procedures can offer and has contributed to the strong focus by the statistical community on model *estimation* to the neglect of the logical predecessor to this step, namely model *identification*. Another culprit underlying this benign neglect is the close historical connection between mathematics and statistics whereby statisticians tend to work on problems where theorems and other analytical solutions are attainable (*e.g.* sampling distributions and asymptotics). Such solutions are necessarily conditional on the underlying model being specified up to a small number of identifiable parameters that summarize the relationship of the predictor variables to the response variable through the first few moments of the conditional distribution, $f(y|x)$. For example the common regression model takes the form:

$$\mu(y|x) = \sum_{j=1}^{J} x_j \beta_j \tag{4}$$

$$\sigma(y|x) = constant \tag{5}$$

The implicit parameter J is not part of the explicit formulation nor is the precise specification of which x_j's define the model for the mean parameter μ. Traditional statistics provides very useful information on the sampling distribution of the estimates $\hat{\beta}_j$ for a fixed set of x_j's but no formalism for saying which x's are needed.

The relatively small effort by the statistical community in model identification has focused on marrying computing horsepower with human judgment (as opposed to fully automated procedures). The general problem is deciding how large or complex a model the available data will support. By directly and explicitly focusing on mean squared prediction error, statisticians have long understood the basic tradeoff between bias (too small a model) and variance (too large a model) in model selection. Algorithms (Furnival and Wilson, 1978) and methods (Mallows, 1973) have been used extensively in identifying candidate models summarized by model accuracy and model size. The primary reason that human judgment is crucial in this process is that algorithmic optimality does not and cannot include qualitative distinctions between competing models of similar size – for example, if the accuracy/availability/cost of the variables differ. So it is largely human expertise that is used to select (or validate) a model, or a few models, from the potentially large pools of candidate models.

The statistician's tendency to avoid complete automation out of respect for the challenges of the data, and the historical emphasis on models with interpretable structure, has led that community to focus on problems with a more manageable number of variables (a dozen, say) and cases (several hundred typically) than may be encountered in KDD problems, which can be orders of magnitude larger at the outset.[4] With increasingly huge and amorphous databases, it is clear that methods for automatically hunting down possible patterns worthy of fuller, interactive attention, are required. The existence of such tools can free one up to, for instance, posit a wider range of candidate data features and basis functions (building blocks) than one would wish to deal with, if one were specifying a model structure "by hand".

This obvious need is gaining sympathy but precious little has resulted. The subsections below highlight some of the areas that further underlie the hesitation of automating model identification by statisticians.

[4]Final models are often of similar complexity; it's the magnitude of the initial candidate set of variables and cases that is usually larger in KDD.

Statistical versus practical significance

A common approach to addressing the complexity and size of *model space* is to limit model growth in the model fitting/learning stage. This is almost always accomplished using a statistical test of significance at each step in the incremental model building stage. Thus for example, one could use a standard χ^2 test of independence between two nominal variables as a means to limit growth of a model that searches for "significant" association. The main problem with this approach is that significance levels depend critically on n, the sample size, such that as n increases, even trivial differences attain statistical significance. Statisticians ameliorate this problem by introducing context to better qualify findings as "significant."

Simpson's paradox

A related problem with automated search procedures is that they can often be completely fooled by anomalous association patterns, even for small datasets. An accessible and easily understood example (Freedman, Pisani, and Purves, 1978) concerns admission to graduate school at UC Berkeley in 1973. Across major departments, 30% of 1835 female applicants were admitted while 44% of 2691 male applicants were admitted. Do these disparate fractions indicate sex bias? On the face yes, but if the applicants and admissions are broken down by department, then the fractions of the two sexes admitted shows a very different story, where one might even argue that "reverse" sex bias is present! The "paradox" here is that the choice of major is *confounded* with sex – namely that females tend to apply to majors that are harder to get into while males apply to "easy" majors.

The implication of this paradox is that KDD tools which attack large databases looking for "interesting" associations between pairs of variables must also contain methods to search for potential confounders. Computationally, this changes the problem from an n^2 to an n^3 operation (or higher if one considers more confounders). The computational burden can only be avoided by providing knowledge about potential confounders to the discovery algorithm. While this is in principle possible, it is unlikely to be sufficient since common sense knowledge often suggests what confounders might be operating. Statisticians have long brought these common sense insights to the problem rather than delegate them to automata.

Selection bias

Automated knowledge discovery systems are applied to databases with the expectation of translating *data* into *information*. The bad news is that often the available data is not representative of the population of interest and the worse news is that the data itself contains no hint that there is a potential bias present. Namely, it's more an issue of what is *not* in a data set rather than what information it contains. For example [5], suppose that the White House Press Secretary is using a KDD (*e.g.* information retrieval) tool to browse through email messages to PRESIDENT@WHITEHOUSE.GOV for those that concern health care reform. Suppose that she finds a 10:1 ratio of pro-reform to anti-reform messages, leading her to assert that "Americans favor reform by a 10:1 ratio" followed by the worrisome rejoinder "and Government can fix it." But it may well be that people dissatisfied with the health care system are *more likely* to "sound off" about their views than those who are satisfied. Thus even if the true distribution of views on health care reform has mean "score" of zero, self-selected samples that are heavily biased toward s one of the tails of this distribution will give a very misleading estimate of the true situation. It is not realistic to expect automated tools to identify such instances. It is probably even less realistic to expect users (*e.g.* lawyers) of such systems to critically question such interesting "facts."

Quantifying Visual Capabilities

Today's data analyst is very dependent on interactive analysis where numerical and graphical summaries are computed or displayed "on the fly". Successful instances of data mining by statisticians are often sprinkled with cries of "aha" whereby some subject matter (context) information, or unexpected behavior in a plot, is discovered in the course of the interaction with the data. This discovery can change the intended course of subsequent analysis steps in quite unpredictable ways. Assuming that it is a very hard problem to include common sense and context information in automated modeling systems, this leaves automated interpretation of plots as a promising area to explore. There are two problems that have served as a barrier to statisticians in this regard:

1. it is hard to quantify a procedure to capture the *unexpected* in plots.
2. even if this could be accomplished, one would need to describe how this maps into the next analysis step in the automated procedure.

What is sorely needed in the statisticians armory is a way to represent meta-knowledge about the problem at hand and about the procedures commonly used. This suggests an opportunity where the KDD and statistical communities can complement their skills and work together to provide an acceptable and powerful solution.

[5] A less modern but more realistic situation occurred in US politics when three major polls overwhelmingly projected Dewey over Truman in the 1948 presidential election — too bad for Dewey (the Republican) that there was a discrepancy between the voting public and those with phone service.

Summary

Numerous powerful analytic techniques, applicable to problems of Knowledge Discovery in Databases, have emerged from the field of statistics in recent decades – especially as the influence of computing has grown over that of mathematics. "Thinking in statistical terms" can be vital to the quality of models induced from data. In particular, the tendency of the statistical community to propagate uncertainty in their models through sampling distributions, their familiarity with the need to regularize models (trade off accuracy and complexity), and their perseverance in checking model assumptions and stability (through residual and graphical analyses) are strengths. KDD researchers can learn from these perspectives in order to improve the performance and stability of their techniques. Similarly, KDD, machine learning, and neural network techniques have gained in popularity partly as a way of "avoiding statistics". Statisticians can learn from this the need to do a better job of communicating their methods, as well as clarifying and streamlining ways of injecting meta-data information into the modeling process.

References

Akaike, H. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In Proceedings of the Second International Symposium on Information Theory, eds. Petrov and Csaki, 267–281, Budapest: Kiado Academy.

Belsley, D. A.; Kuh, E.; and Welsch, R. E. 1980. *Regression Diagnostics: Identifying influential data and sources of collinearity.* New York: John Wiley & Sons.

Bishop, Y. M. M.; Fienberg, S. E.; and Holland, P. W. 1975. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Massachusetts: MIT Press.

Breiman, L. 1994b. Stacked Regressions, Technical Report 367, Dept. Statistics, UC Berkeley.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees.* Monterey, California: Wadsworth & Brooks.

Cleveland, W. S. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical. Association* 74: 829–836.

Dempster, A. P.; Laird, N. M.; and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* 39: 1–38.

Efron, B.; and Gong, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician* 37: 36–48.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7: 1–26.

Faraway, J. J. 1991. On the cost of Data Analysis, Technical Report 199, Dept. Statistics, Univ. Michigan, Ann Arbor.

Fayyad, U. M.; Piatetsky-Shapiro, G; Smyth, P; and Uthurusamy, R. 1995 (in press). *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press.

Freedman, D.; Pisani, R.; and Purves, R. 1978. *Statistics.* New York: WW Norton & Co.

Friedman, J. H. 1991. Multiple Adaptive Regression Splines (with discussion). *Annals of Statistics* 19: 1–141.

Friedman, J. H.; and Tukey, J. W. 1974. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers* 23: 881–889.

Furnival, G. M.; and Wilson, R. W. 1974. Regression by leaps and bounds. *Technometrics* 16: 499–511.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 62: 1179–1186.

Hastie, T.; and Tibshirani, R. 1990. *Generalized Additive Models.* London: Chapman & Hall.

Huber, P. J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35: 73–101.

Mallows, C. L. 1973. Some Comments on Cp. *Technometrics* 15: 661–675.

McCullagh, P.; and Nelder, J. A. 1989. *Generalized Linear Models (2nd Ed.)* London: Chapman & Hall.

Mosteller, F.; and Tukey, J. W. 1977. *Data Analysis and Regression.* Reading Massachusetts: Addison-Wesley.

Nelder, J. A.; and Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society A* 135: 370–384.

O'Sullivan, F.; Yandell, B. S.; and Raynor, W. J. Jr. 1986. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* 81: 96–103.

Rissanen, J. 1978. Modeling by Shortest Data Description. *Automatica* 14: 465–471.

Schreuder, H. T. 1986. Quenouille's estimator. *Encyclopedia of Statistical Science* 7: 473–476. New York: John Wiley & Sons.

Tukey, J. W. 1977. *Exploratory Data Analysis.* Reading, Massachusetts: Addison-Wesley.

Wolpert, D. 1992. Stacked Generalization. *Neural Networks* 5: 241–259.