IN SUMMARY

Algorithmic modeling is a very important area of statistics. It has evolved naturally in environments with lots of data and lots of decisions. But you can do it without suffering the Occam dilemma; for example, use medium trees with interpretable GAMs in the leaves. They are very accurate and interpretable. And you can do it with data modeling tools as long as you (i) ignore most textbook advice, (ii) embrace the blessing of dimensionality, (iii) use constraints in the fitting optimizations (iv) use regularization, and (v) validate the results.

Comment

Emanuel Parzen

1. BREIMAN DESERVES OUR APPRECIATION

I strongly support the view that statisticians must face the crisis of the difficulties in their practice of regression. Breiman alerts us to systematic blunders (leading to wrong conclusions) that have been committed applying current statistical practice of data modeling. In the spirit of "statistician, avoid doing harm" I propose that the first goal of statistical ethics should be to guarantee to our clients that any mistakes in our analysis are unlike any mistakes that statisticians have made before.

The two goals in analyzing data which Leo calls prediction and information I prefer to describe as "management" and "science." Management seeks *profit*, practical answers (predictions) useful for decision making in the short run. Science seeks *truth*, fundamental knowledge about nature which provides understanding and control in the long run. As a historical note, Student's t-test has many scientific applications but was invented by Student as a management tool to make Guinness beer better (bitter?).

Breiman does an excellent job of presenting the case that the practice of statistical science, using only the conventional data modeling culture, needs reform. He deserves much thanks for alerting us to the algorithmic modeling culture. Breiman warns us that "if the model is a poor emulation of nature, the conclusions may be wrong." This situation, which I call "the right answer to the wrong question," is called by statisticians "the error of the third kind." Engineers at M.I.T. define "suboptimization" as "elegantly solving the wrong problem."

Emanuel Parzen is Distinguished Professor, Department of Statistics, Texas A&M University, 415 C Block Building, College Station, Texas 77843 (e-mail: eparzen@stat.tamu.edu).

Breiman presents the potential benefits of algorithmic models (better predictive accuracy thandata models, and consequently better information about the underlying mechanism and avoiding questionable conclusions which results from weak predictive accuracy) and support vector machines (which provide almost perfect separation and discrimination between two classes by increasing the dimension of the feature set). He convinces me that the methods of algorithmic modeling are important contributions to the tool kit of statisticians.

If the profession of statistics is to remain healthy, and not limit its research opportunities, statisticians must learn about the cultures in which Breiman works, *but also* about many other cultures of statistics.

2. HYPOTHESES TO TEST TO AVOID BLUNDERS OF STATISTICAL MODELING

Breiman deserves our appreciation for pointing out generic deviations from standard assumptions (which I call bivariate dependence and two-sample conditional clustering) for which we should routinely check. "Test null hypothesis" can be a useful algorithmic concept if we use tests that diagnose in a model-free way the directions of deviation from the null hypothesis model.

Bivariate dependence (correlation) may exist between features [independent (input) variables] in a regression causing them to be proxies for each other and our models to be unstable with different forms of regression models being equally well fitting. We need tools to routinely test the hypothesis of statistical independence of the distributions of independent (input) variables.

Two sample conditional clustering arises in the distributions of independent (input) variables to discriminate between two classes, which we call the conditional distribution of input variables X given each class. Class I may have only one mode (cluster) at low values of X while class II has two modes

(clusters) at low and high values of X. We would like to conclude that high values of X are observed only for members of class II but low values of X occur for members of both classes. The hypothesis we propose testing is equality of the pooled distribution of both samples and the conditional distribution of sample I, which is equivalent to $P[class I|X] = P[class I]$. For successful discrimination one seeks to increase the number (dimension) of inputs (features) X to make $P[class I|X]$ close to 1 or 0.

3. STATISTICAL MODELING, MANY CULTURES, STATISTICAL METHODS MINING

Breiman speaks of two cultures of statistics; I believe statistics has *many cultures*. At specialized workshops (on maximum entropy methods or robust methods or Bayesian methods or ...) a main topic of conversation is "Why don't all statisticians think like us?"

I have my own eclectic philosophy of statistical modeling to which I would like to attract serious attention. I call it "statistical methods mining" which seeks to provide a framework to synthesize and apply the past half-century of methodological progress in computationally intensive methods for statistical modeling, including EDA (exploratory data analysis), FDA (functional data analysis), density estimation, Model DA (model selection criteria data analysis), Bayesian priors on function space, continuous parameter regression analysis and reproducing kernels, fast algorithms, Kalman filtering, complexity, information, quantile data analysis, nonparametric regression, conditional quantiles.

I believe "data mining" is a special case of "data modeling." We should teach in our introductory courses that one meaning of statistics is "statistical data modeling done in a systematic way" by an iterated series of stages which can be abbreviated SIEVE (specify problem and general form of models, identify tentatively numbers of parameters and specialized models, estimate parameters, validate goodness-of-fit of estimated models, estimate final model nonparametrically or algorithmically). MacKay and Oldford (2000) brilliantly present the statistical method as a series of stages PPDAC (problem, plan, data, analysis, conclusions).

4. QUANTILE CULTURE, ALGORITHMIC MODELS

A culture of statistical data modeling based on quantile functions, initiated in Parzen (1979), has been my main research interest since 1976. In my discussion to Stone (1977) I outlined a novel approach to estimation of conditional quantile functions which I only recently fully implemented. I would like to extend the concept of algorithmic statistical models in two ways: (1) to mean data fitting by representations which use approximation theory and numerical analysis; (2) to use the notation of probability to describe empirical distributions of samples (data sets) which are not assumed to be generated by a random mechanism.

My quantile culture has not yet become widely applied because "you cannot give away a good idea, you have to sell it" (by integrating it in computer programs usable by applied statisticians and thus promote statistical methods mining).

A quantile function $Q(u)$, $0 \le u \le 1$, is the inverse $F^{-1}(u)$ of a distribution function $F(x)$, $-\infty < x < \infty$. Its rigorous definition is $Q(u) = \inf(x: F(x) \ge u)$. When F is continuous with density f, $F(Q(u)) = u$, $q(u) = Q'(u) = 1/f(Q(u))$. We use the notation Q for a true unknown quantile function, $\tilde{Q}$ for a raw estimator from a sample, and $\hat{Q}$ for a smooth estimator of the true Q.

Concepts defined for $Q(u)$ can be defined also for other versions of quantile functions. Quantile functions can "compress data" by a five-number summary, values of $Q(u)$ at $u = 0.5, 0.25, 0.75, 0.1, 0.9$ (or $0.05, 0.95$). Measures of location and scale are $QM = 0.5(Q(0.25) + Q(0.75))$, $QD = 2(Q(0.75) - Q(0.25))$. To use quantile functions to identify distributions fitting data we propose the quantile—quartile function $Q/Q(u) = (Q(u) - QM)/QD$. Five-number summary of distribution becomes $QM, QD, Q/Q(0.5)$ skewness, $Q/Q(0.1)$ left-tail, $Q/Q(0.9)$ right-tail. Elegance of $Q/Q(u)$ is its universal values at $u = 0.25, 0.75$. Values $|Q/Q(u)| > 1$ are outliers as defined by Tukey EDA.

For the fundamental problem of comparison of two distributions F and G we define the comparison distribution $D(u; F, G)$ and comparison density $d(u; F, G) = D'(u; F, G)$. For F, G continuous, define $D(u; F, G) = G(F^{-1}(u))$, $d(u; F, G) = g(F^{-1}(u))/f(F^{-1}(u))$ assuming $f(x) = 0$ implies $g(x) = 0$, written $G \ll F$. For F, G discrete with probability mass functions p_F and p_G define (assuming $G \ll F$) $d(u; F, G) = p_G(F^{-1}(u))/p_F(F^{-1}(u))$.

Our applications of comparison distributions often assume F to be an unconditional distribution and G a conditional distribution. To analyze bivariate data (X, Y) a fundamental tool is *dependence* density $d(t, u) = d(u; F_Y, F_{Y|X=Q_X(t)})$. When X, Y is jointly continuous,

$$d(t, u) = f_{X,Y}(Q_X(t), Q_Y(u))/f_X(Q_X(t))f_Y(Q_Y(u)).$$

The statistical independence hypothesis $F_{X,Y} = F_X F_Y$ is equivalent to $d(t, u) = 1$, all t, u. A fundamental formula for estimation of conditional quantile functions is

$$Q_{Y|X=x}(u) = Q_Y(D^{-1}(u; F_Y, F_{Y|X=x})) = Q_Y(s), u = D(s; F_Y, F_{Y|X=x}).$$

To compare the distributions of two univariate samples, let Y denote the continuous response variable and X be binary 0, 1 denoting the population from which Y is observed. The comparison density is defined (note F_Y is the pooled distribution function)

$$d_1(u) = d(u; F_Y, F_{Y|X=1}) = P[X = 1|Y = Q_Y(u)]/P[X = 1].$$

5. QUANTILE IDEAS FOR HIGH DIMENSIONAL DATA ANALYSIS

By high dimensional data we mean multivariate data $(Y_1, \ldots, Y_m)$. We form approximate high dimensional comparison densities $d(u_1, \ldots, u_m)$ to test statistical independence of the variables and, when we have two samples, $d_1(u_1, \ldots, u_m)$ to test equality of sample I with pooled sample. All our distributions are empirical distributions but we use notation for true distributions in our formulas. Note that

$$\int_0^1 du_1 \ldots \int_0^1 du_m d(u_1, \ldots, u_m) d_1(u_1, \ldots, u_m) = 1.$$

A decile quantile bin $B(k_1, \ldots, k_m)$ is defined to be the set of observations $(Y_1, \ldots, Y_m)$ satisfying, for $j = 1, \ldots, m$, $Q_{Y_j}((k_j - 1)/10) < Y_j \leq Q_{Y_j}(k_j/10)$. Instead of deciles $k/10$ we could use k/M for another base M.

To test the hypothesis that $Y_1, \ldots, Y_m$ are statistically independent we form for all $k_j = 1, \ldots, 10$,

$$d(k_1, \ldots, k_m) = P[\text{Bin}(k_1, \ldots, k_m)]/P[\text{Bin}(k_1, \ldots, k_m)|\text{independence}].$$

To test equality of distribution of a sample from population I and the pooled sample we form

$$d_1(k_1, \ldots, k_m) = P[\text{Bin}(k_1, \ldots, k_m)|\text{ population } I]/P[\text{Bin}(k_1, \ldots, k_m)|\text{ pooled sample}]$$

for all $(k_1, \ldots, k_m)$ such that the denominator is positive and otherwise defined arbitrarily. One can show (letting X denote the population observed)

$$d_1(k_1, \ldots, k_m) = P[X = I|\text{ observation from Bin}(k_1, \ldots, k_m)]/P[X = I].$$

To test the null hypotheses in ways that detect directions of deviations from the null hypothesis our recommended first step is *quantile data analysis* of the values $d(k_1, \ldots, k_m)$ and $d_1(k_1, \ldots, k_m)$.

I appreciate this opportunity to bring to the attention of researchers on high dimensional data analysis the potential of quantile methods. My conclusion is that statistical science has many cultures and statisticians will be more successful when they emulate Leo Breiman and apply as many cultures as possible (which I call statistical methods mining). Additional references are on my web site at stat.tamu.edu.

Rejoinder

Leo Breiman

I thank the discussants. I'm fortunate to have comments from a group of experienced and creative statisticians—even more so in that their comments are diverse. Manny Parzen and Bruce Hoadley are more or less in agreement, Brad Efron has serious reservations and D. R. Cox is in downright disagreement.

I address Professor Cox's comments first, since our disagreement is crucial.

D. R. COX

Professor Cox is a worthy and thoughtful adversary. We walk down part of the trail together and then sharply diverge. To begin, I quote: "Professor Breiman takes data as his starting point. I would prefer to start with an issue, a question, or a scientific hypothesis,...." I agree, but would expand the starting list to include the prediction of future events. I have never worked on a project that has started with "Here is a lot of data; let's look at it and see if we can get some ideas about how to use it." The data has been put together and analyzed starting with an objective.

C1 Data Models Can Be Useful

Professor Cox is committed to the use of data models. I readily acknowledge that there are situations