Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

In refereeing papers for NIPS I was struck by the growing emphasis on mathematical theory. Having lived for forty years in a field plagued with theory, and beginning life as a probability theorist, I thought I might make a few remarks that summarize what experience I have with theory as it may be relevant to machine learning.

1. WHAT IS THEORY?

This may be difficult to define. A rough definition is:

THEORY = mathematical framework plus theorems plus proofs.
"No theorems" implies "no theory."

One problem in the field of statistics has been that everyone wants to be a theorist. Part of this is envy—the real sciences are based on mathematical theory. In the universities for this century, the glamor and prestige has been in mathematical models and theorems, no matter how irrelevant.

The Mathematics of Generalization, Ed. David Wolpert, SFI Studies in
the Sciences of Complexity, Proc. Vol XX, Addison-Wesley, 1995

As a result of the would-be mathematicians in statistics, it has been dominated by useless theory and fads.

- Decision Theory
- Asymptotics
- Robustness
- Nonparametric One and Two Sample Tests
- One-Dimensional Density Estimation
- Etc.

If statistics is an applied field and not a minor branch of mathematics, then more than 99% of the published papers are useless exercises. (The other colleagues in statistics I have spoken to say this is an exaggeration and peg the percentage at 95%. Either way it is significant). The result is a downgrading of sensibility and intelligence.

But among all of the trash, there are a few places where theory has been useful. To understand the potential usefulness of theory, I look at this a bit more.

2. USES OF THEORY

- **Comfort**: We knew it worked, but it's nice to have a proof.
- **Insight**: Aha! So that's why it works.
- **Innovation**: At last, a mathematically proven idea that applies to data.
- **Suggestion**: Something like this might work with data.

3. EXAMPLES: POST WORLD WAR II

3.1 COMFORT

Mainly asymptotics.

1. Gordon and Olshen proved CART consistent; i.e., as the sample size goes to infinity, the CART risk converges to the Bayes risk. The estimated sample size for the proof to take force is in the neighborhood of a million.
2. Cover proves that as the sample size becomes infinite, the nearest-neighbor risk becomes less than twice the Bayes risk (also provides insight).
3. Many theorems that say: given a specified class of basis functions, linear combinations of these are dense in some large class of smooth functions; i.e., any sufficiently smooth function can be arbitrarily closely approximated by a linear combination of the basis functions.

4. Charles Stone proves that if data $\{(y_n, \mathbf{x}_n), n = l, \ldots, N\}$ where $\mathbf{x}_n$ is a vector in M dimensions, is sampled from $y = f(\mathbf{x}) + \varepsilon$, and some procedure used to estimate f, then there is always a continuous differentiable f such that the root mean squared error in the estimate is at least order of $N^{(-2/(2+M))}$. Thus, no procedure can hope to approximate all continuous differentiable functions in a high-dimensional space without the use of very large sample sizes.

Stone also proved later that if f is known to be in the class of continuous differentiable functions consisting of sums of functions of $J < M$ variables, then there is an approximation method such that the root mean squared error is uniformly of order $N^{(-2/(2+J))}$. These two results also provide insight.

But sometimes comfort results can be misleading. For instance, things that hold for nearly infinite sample size blow up for finite sample sizes. One example is the use of Akaike's penalty to select the dimension of a model. Its asymptotic properties simply do not hold for moderate sample sizes and its use gives poor results.

3.2 INSIGHT

1. Donoho's recent work on image processing under sparsity constraints.
2. Recent work on tomography.

Note that in both cases, physical laws made possible more precise modeling of data.

3.3 SUGGESTION

Here it is difficult to cite examples because the suggestive effect of some piece of theory on applied results in a different area are often undocumented. In my own work:

1. The pi-method, for approximating functions using noisy data, was suggested by results in mathematical approximation theory.
2. The ACE algorithm is a data implementation of results suggested by random variable inequalities.

3.4 INNOVATION

1. The theoretical work by Weiner and others on the spectral analysis of stationary time series penetrated statistics following Tukey's heuristic work on estimation of the spectrum. This opened up the field of time series analysis.
2. Shannon's work on information theory led to some important early work by statisticians, but the applications and further work has passed onto the engineering fields.

3. Efron's invention of the bootstrap and his early work on its asymptotic properties established its credentials and it is now extensively used in applied work. But even so, the analytics of its performance are known only in simple cases.

The above list of useful theories was not meant to be inclusive, but even a more inclusive list would be very short. A possible reason is that it is difficult to formulate reasonable analytic models for complex data.

Notice also that none of the useful theory on the list were of the Grand Unification Type theory. Following WW II there was an effort to provide GUT theory for statistics in the form of decision theory, and works that hung off of this framework. In spite of intense activity, none of this work has had any effect on the day-to-day practice of statistics, or even on present-day theory. It slumbers in its own sanctified graveyard.

Mathematical theory is not critical to the development of machine learning.

But scientific inquiry is.

3.5 INQUIRY

INQUIRY = sensible and intelligent efforts to understand what is going on. For example:

- mathematical heuristics
- simplified analogies (like the Ising Model)
- simulations
- comparisons of methodologies
- devising new tools
- theorems where useful (rare!)
- shunning panaceas

Regarding this last point, every field gets frozen when a certain tool becomes a panacea (a.k.a. fad). For instance, for years in statistics, everything had to be robust. In machine learning, the current panacea is a sigmoid network fitted using backpropagation.

My colleague, Jerry Friedman, once told me an old folk saying: "Give a man a hammer and every problem looks like a nail." What is needed is not one hammer, but many different tools along with a sense of which ones to use.

For example, I have image data of healthy and diseased bones that was fit by use of neural networks at Mayo Clinic and produced 95% test set accuracy. Crude CART runs got 96% accuracy. I think that any reasonable classification method used on this data would produce comparable accuracy, possibly even linear discriminant analysis.

Our fields would be better off with far fewer theorems, less emphasis on faddish stuff, and much more scientific inquiry and engineering. But the latter requires real thinking.

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?

It makes research more interesting to know that there is no one universally best method. What is best is data dependent. Sometimes "least glamorous" methods such as nearest neighbor are best. We need to learn more about what works best where. But emphasis on theory often distracts us from doing good engineering and living with the data.