

Comments on “Hedging Predictions in Machine Learning,” by Gammerman and Vovk

Glenn Shafer
Royal Holloway and Rutgers University
June 12, 2006

This article provides an excellent explanation of the fundamentals of conformal prediction. I have already begun recommending it to those who want to master the method without wading into the more comprehensive and intricate exposition in \cite{vovk/gammerman/shafer:2005}.

Like all good ideas, conformal prediction has a complex ancestry. As Gammerman and Vovk explain, they invented the method as a result of their study of work by Chervonenkis, Vapnik, Kolmogorov, and Martin-Löf. But they subsequently discovered related ideas in earlier work by mathematical statisticians. As we explain on pp. 256-257 of \cite{vovk/gammerman/shafer:2005}, Sam Wilks, Abraham Wald, and John Tukey developed non-parametric tolerance regions based on permutation arguments in the 1940s, and Donald Fraser and J. H. B. Kemperman used the same idea to construct prediction regions in the 1950s. From our viewpoint, Fraser and Kemperman were doing conformal prediction in the special case where y s are predicted without the use of x s. It is easy (once you see it) to extend the method to the case where x s are used, and Kai Takeuchi has told us that he explained this in the early 1970s, first in lectures at Stanford and then in a book that appeared in Japanese in 1975 \cite{takeuchi:1975}. Takeuchi's idea was not taken up by others, however, and the rediscovery, thorough analysis, and extensions by Gammerman and Vovk are remarkable achievements.

Because it brings together methods well known to mathematical statisticians (permutation methods in non-parametrics) and a topic now central to machine learning (statistical learning theory), the article prompts me to ask how these two communities can be further unified. How can we make sure the next generation of mathematical statisticians and computer scientists will have full access to each other's experience and traditions?

Statistical learning theory is limited in one very important respect: it considers only the case where examples are independent and identically distributed, or at least exchangeable. The iid case has also been central to statistics ever since Jacob Bernoulli proved the law of large numbers at the end of the 17th century, but its inadequacy was always obvious. Leibniz made the point in his letters to Bernoulli: the world is in constant flux; causes do not remain constant, and so probabilities do not remain constant. Perhaps Leibniz's point is a counterexample to itself, for it is as topical in 2006 as it was in the 1690s. In the most recent issue of *Statistical Science*, David Hand gives as one of his reasons for skepticism about apparent progress in classifier technology the fact that “in many, perhaps most, real classification problems the data points in the design set are not, in fact, randomly drawn from the same distribution as the data points to which the classifier will be applied” \cite{hand:2006}.

It is revealing that Hand finds it necessary to say this three centuries after Leibniz. We can cite methods that have been developed to deal with non-iid data:

1. Starting at the end of the 18th century, probabilists used models in the y s are independent only given the x s. To get results, they then made strong assumptions about the distribution of the y s. If we assume the y s are Gaussian with constant variance and means linear in the x s, we get the Gauss linear model, so-called because Gauss used it to prove the optimality of least squares vovk/nouretdinov/gammerman:2006}.
2. Starting with Markov at the end of the 19th century, probabilists have studied stochastic process models – probability models for successive examples that are not necessarily iid.
3. Statisticians often take differences between successive observations, perhaps even higher-order differences, in attempt to get something that looks iid.
4. A major topic in machine learning, prediction with expert advice, avoids making any probability assumptions at all. Instead, one specifies a class of prediction procedures that one is competing with \cite{cesa-bianchi/lugosi:2006}.

But we have stayed so true to Bernoulli in our overview of what statistics is about that we seldom ask potential statisticians and data analysts to look at a list like this. A general course in statistical inference usually still studies the iid case, leaving each alternative to be taken up as something distinct, often in some specialized discipline, such as psychometrics, econometrics, or machine learning, whose special terminology makes its results inaccessible to others. Except perhaps in a course in “consulting,” we seldom ponder or teach how to compare and choose among the alternatives.

Reinforcing the centrality of the iid picture is the centrality of the Cartesian product as the central structure for relational databases. Neither in statistics nor in computer science have we built on Art Dempster’s now classic (but unfortunately not seminal) article on alternatives to the Cartesian product as a data structure \cite{dempster:1971}.

More than 15 years ago I urged that statistics departments embrace the insights of specialized disciplines such as econometrics and machine learning in order to regain the unifying educational role that they held in the mid-twentieth century \cite{shafer:1990}. It is now clear that this will not happen. Statistics is genetically imprinted with the Bernoulli code \cite{bru:2006}. Perhaps the machine learning community, which has had the imagination to break out of the probabilistic mode altogether with its concept of prediction with expert advice, should pick up this leadership mantle.

cesa-bianchi/lugosi:2006

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*, Cambridge University Press, 2006.

bru:2006

Bernard Bru. The Bernoulli code. *Electronic Journal for History of Probability and Statistics* (www.jehps.net). Vol. 2, No. 1, June 2006.

dempster:1971

A. P. Dempster. An overview of multivariate data analysis. *Journal of Multivariate Analysis*, Volume 1, pages 316-346. 1971.

hand:2006

David J. Hand. Classifier technology and the illusion of progress (with discussion). *Statistical Science* Volume 21, pages 1-14. 2006.

shafer:1990

Glenn Shafer. The unity and diversity of probability (with discussion). *Statistical Science* Volume 5, pages 435-462. 1990.

takeuchi:1975

Reference 16 of the paper

vovk/gammerman/shafer:2005

Reference 23 of the paper