Abstracta SPECIAL ISSUE III, pp. 10 – 17, 2009

COMMENTS ON HARMAN AND KULKARNI'S "RELIABLE REASONING"

Glenn Shafer

Gil Harman and Sanjeev Kulkarni have written an enjoyable and informative book that makes Vladimir Vapnik's ideas accessible to a wide audience and explores their relevance to the philosophy of induction and reliable reasoning. The undertaking is important, and the execution is laudable.

Vapnik's work with Alexey Chervonenkis on statistical classification, carried out in the Soviet Union in the 1960s and 1970s, became popular in computer science in the 1990s, partly as the result of Vapnik's books in English. Vapnik's statistical learning theory and the statistical methods he calls support vector machines now dominate machine learning, the branch of computer science concerned with statistical prediction, and recently (largely after Harman and Kulkarni completed their book) these ideas have also become well known among mathematical statisticians.

A century ago, when the academic world was smaller and less specialized, philosophers, mathematicians, and scientists interested in probability, induction, and scientific methodology talked with each other more than they do now. Keynes studied Bortkiewicz, Kolmogorov studied von Mises, Le Dantec debated Borel, and Fisher debated Jeffreys. Today, debate about probability and induction is mostly conducted within more homogeneous circles, intellectual communities that sometimes cross the boundaries of academic disciplines but overlap less in their discourse than in their membership. Philosophy of science, cognitive science, and machine learning are three of these communities. The greatest virtue of this book is that it makes ideas from these three communities confront each other. In particular, it looks at how Vapnik's ideas in machine learning can answer or dissolve questions and puzzles that have been posed by philosophers.

This leaves out, of course, many other communities that debate probability and reliable reasoning: mathematical probabilists, the many tribes of mathematical statisticians,

economists, psychologists, information theorists, and even other tribes of computer scientists, including those within machine learning who study prediction with expert advice [2]. The work of Vapnik and Chervonenkis is only a tiny part of the vast literature on probability and prediction that is relevant to philosophy's questions about induction and reliable reasoning, and the next step beyond Harman and Kulkarni's book is surely to try to fit what they have done into a larger picture.

From a historical viewpoint, and also from the viewpoint of modern mathematical probability, Vapnik's statistical learning theory makes a very special assumption: it assumes that repeated observations are drawn from the same probability distribution. As Harman and Kulkarni explain (p. 35), "we assume that the data represent a random sample arising from the background probability distribution, and we assume that new cases that are encountered are also randomly produced by that distribution." This is the famous assumption that observations are independently and identically distributed. It can be weakened slightly to the assumption that the observations are exchangeable – i.e., that their probability distribution does not change when the order is permuted. Are there really many applications where either assumption is reasonable?

Leibniz thought not. In 1703, Jacob Bernoulli wrote to Leibniz to explain how his law of large numbers would use past examples to find probabilities for future ones: "For example, if I perceive, having made the experiment in very many pairs of young and old, that it happens 1000 times that the young person outlives the old person and the reverse happens only 500 times, then I may safely enough conclude that it is twice as probable that a young person will outlive an old one as the reverse." Leibniz responded skeptically: "Who is to say that the following result will not diverge somewhat from the law of all the preceding ones – because of the mutability of things? New diseases attack mankind. Even if you have observed the results for any number of deaths, you have not therefore set limits on the nature of things so that they cannot vary in the future." (See [1], pp.38-39.)

The history of mathematical probability in the three centuries after Leibniz's exchange with Bernoulli can be framed as a continuation of their debate. Mathematicians continually refined Bernoulli's law of large numbers, but its success in applications was spotty. In the 19th century, Laplace's theory of errors of measurement reigned in astronomy,

while its applications in human affairs were rightly ridiculed. Frank Knight, founder of the Chicago school of economics, coined the distinction between "risk" and "uncertainty" to distinguish between the situation of an insurance company, which can count on the law of large numbers, and the situation of a businessman, who does not enjoy the luxury of many repeated chances under constant conditions.

The great accomplishment of mathematical probability during the twentieth century was to move beyond the picture of successive independent draws from a single probability distribution to the idea of a stochastic process, in which probabilities evolve. This change was already underway in the 1920s, with the explosion of work on Markov chains [3] and Weiner's application of functional analysis to model Brownian motion. It was consecrated in 1953 by Joe Doob's general framework for stochastic processes, which applied to continuous as well as discrete time [4]. By 1960, Jerzy Neyman could declare that science had become the study of stochastic processes [6].

Neyman saw four periods in the history of indeterminism in science:

1. *Marginal indeterminism*, the period in the 19th century when scientific research was indeterministic except in the domain of errors of measurement.

2. *Static indeterminism*, the period at the end of the 19th and beginning of the 20th century when populations were the main subject of scientific study, so that the idea of independent draws from populations was dominant.

3. *Static indeterministic experimentation*, the period from 1920 to 1940 when R. A. Fisher's ideas were dominant and the basic ideas of statistical testing and estimation were developed.

4. *Dynamic indeterminism*, already in full swing in 1960, when every serious study in science was a study of some evolutionary chance mechanism.

"In order that the applied statistician be in a position to cooperate effectively with the modern experimental scientist," Neyman declared, "the theoretical equipment of the statistician must include familiarity and capability of dealing with stochastic processes."

In the half century since Neyman wrote, the theory and applications of stochastic processes have developed as he envisioned. Natural science and economics are awash with dynamic stochastic modeling. How can it be, then, that Vapnik's work, based on the tired old idea of independent identically distributed observations, has suddenly emerged as so powerful, finding so many applications in biology and other data-rich domains?

Are these domains in which probabilities do not evolve? I doubt it. The data sets that people in machine learning use to test competing methods generally fails tests of exchangeability, so much so that it is standard practice to permute the order of the observations in these data sets before applying methods, such as support vector machines, that assume exchangeability.

In fact, the results of statistical learning theory that use exchangeability – the guarantees of accuracy based on finite Vapnik-Chervonenkis dimension, for example – are so asymptotic (require such cosmic sample sizes in order to give interesting bounds) that they have little to say about the success of support vector machines. (Concerning accurate confidence levels for successive predictions of a support vector machine or other prediction method when exchangeability does hold, see [8,9].)

The key to the success of support vector machines seems to lie elsewhere – in a feature of their implementation that Harman and Kulkarni mention on pp. 85-87: the mapping of data to higher dimensional spaces where classes can be more nearly linearly separated. This mapping is actually implemented implicitly with kernels, which assign to pairs of vectors in the original space the angles between the vectors to which they would be mapped if the mapping were spelled out. Such kernels were studied by probabilists and mathematical statisticians starting in the 1940s, but it was computer scientists implementing support vector machines who first took advantage of them on a large scale, to process the type of data that has now become so common in medicine and other branches of biology, where the number of individuals measured many be reasonable but an immense number of variables are measured on each individual.

Kernels are becoming increasingly important in computer science and mathematical statistics, not only in support vector machines but in other techniques as well. What is crucial in all cases is the choice of the kernel. Choosing the kernel means choosing what features of the observations we want to use for prediction. Choosing which measurements or which aspects of the measurement (the mapping the kernel represents is a mapping from the original measurements to their many aspects) has always been the central question for statistical prediction, and it becomes only more acute in the high-dimensional problems where support vector machines and other kernel techniques are so useful. If I were to fault Harman and Kulkarni on one point, it is that they do not dwell on the experimentation and reasoning that goes into choosing the kernel. This seems to be where applications of machine learning generate new knowledge, and we might learn something from a philosophical analysis. Is the choice of a kernel an example of induction? Is it inference to the best explanation?

One reason support vector machines can be successful in spite of the failure of the exchangeability that Vapnik assumes in all his theoretical work is that the machines rely not so much on stability of the probability distribution from which examples are drawn as on the stability of the relation between the features used for prediction and what is predicted. In order to make this point as clearly as possible, let us write x for the vector of measurements we use for prediction (the *object*) and y for what we predict (the object's *label*). An *example* is a pair (x,y). A kernel is a function K that assigns a real number to every pair of examples (x,y) and (x',y'). We observe n examples, say $(x_1,y_1),\ldots,(x_n,y_n)$ and a new object x_{n+1}, and we want to predict the label y_{n+1}. The support vector machine determined by a kernel K is a way of making this prediction. Exchangeability requires that the n+1 examples $(x_1,y_1),\ldots,(x_n,y_n),(x_{n+1},y_{n+1})$ all be drawn from the same probability distribution. In particular, the x_i should all be drawn from the same distribution. But many methods of prediction, including support vector machines, can do a good job even when the distribution from which the x_i are drawn varies, provided the dependence of y_i on x_i remains somewhat stable. It is enough, for example, if the conditional probabilities $P(y_i|x_i)$ do not change and y_i is independent, given x_i, of the earlier examples [10].

I would also like to add a thought to Harman and Kulkarni's discussion of the contrast between induction and transduction (pp. 90-94). As Vovk, Gammerman, and I argue in [9], the contrast may be clarified if we first discuss *on-line prediction*. When we talk about induction, we usually think about deriving a rule from a batch of examples, say $(x_1,y_1),\ldots,(x_n,y_n)$, and then using that rule for prediction in many future examples. But in an on-line setting, where we see example after example and predict y from x each time, it may be practical to update the prediction rule each time. We predict y_{n+1} from x_{n+1} using a rule we learn from analyzing $(x_1,y_1),\ldots,(x_n,y_n)$, but then we observe y_{n+1}, and so before predicting y_{n+2} from x_{n+2}, we get a new prediction rule by analyzing all n+1 examples $(x_1,y_1),\ldots,(x_n,y_n),(x_{n+1},y_{n+1})$. And so on. We use each rule only once. Is a rule that we use only once a rule? Is finding a rule that we immediately discard induction? The question is obviously relevant to cognitive science if some mental routines are slightly modified whenever they are used.

Finally, I would like to mention recent work on on-line prediction by Vovk, Takemura, and myself [7], which leads to some new insights into the question of whether examples need to be drawn from a probability distribution in order for good probabilistic prediction to be possible. In general, probability predictions are considered good if they pass statistical tests that compare them with what actually happens. Consider, for example, a forecaster who uses information x_i to give a probability p_i. for whether it will rain ($y_i = 1$) or not ($y_i = 0$). It is easy to construct statistical tests for whether the p_i agree with the y_i well enough, and these tests can be reframed as strategies for a gambler who tries to multiply the capital he risks by a large factor betting at the odds given by the p_i. It turns out that the gambler can combine these strategies into a single strategy, which involves a kernel that measures how much a new example (x_{n+1},y_{n+1}) is like old examples $(x_1,y_1),\ldots,(x_n,y_n)$. It also turns out that the forecaster can defeat such a strategy, regardless of how the weather turns out. We call this *defensive forecasting*.

The possibility of defensive forecasting means that good on-line prediction does not depend on examples being drawn from a background probability distribution. The most crucial question in prediction is not whether examples are being chosen from probabilities but whether the prediction is on-line. If the prediction is on-line, there are many ways it can

be done well. Some of these seem to involve the estimation of a background probability distribution, but this is illusory, for the estimate of the background probability distribution can change drastically as prediction proceeds. The important point is that no matter how reality actually chooses the y_i, you can give p_i that avoid extending any trends that might lead to statistical rejection of your forecasting.

In their discussion of reflective equilibrium (pp. 13-19), Harman and Kulkarni mention the situation of a juror, who is scarcely in an on-line setting. The opposing counsels will propose to the juror quite different sequences of examples in which the case at hand might be placed. How to choose? This is Reichenbach's problem of choosing a reference class. Philosophy has something to say here. Bayesians and non-Bayesian theories of subjective probability have something to say. Methods of machine learning, it seems, do not.

Glenn Shafer

Rutgers Business School

&

Department of Computer Science,
Royal Holloway, University of London

gshafer@rutgers.edu

References

[1] Bernoulli, J. (2006). *The Art of Conjecturing*. Translated with an Introduction and Notes by Edith Dudley Sylla. Johns Hopkins: Baltimore.

[2] Cesa-Bianchi, N. & Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press: Cambridge.

[3] Bru, B. (2006). Souvenirs de Bologne. *Journal de la Société française de Statistique,* 144:135-226.

[4] Doob, J.L. (1953). *Stochastic Processes*. Wiley: New York.

[5] Knight, F.H. (1921). *Risk, Uncertainty and Profit*. Houghton Mifflin: Boston.

[6] Neyman, J. (1960). Indeterminism in Science and New Demands on Statisticians. *Journal of the American Statistical Association*, 55:625–639.

[7] Shafer, G. (2008). Defensive Forecasting: How to Use Similarity to Make Forecasts that Pass Statistical Tests. Pp. 215-147 of *Preferences and Similarities*, edited by Giacomo Della Riccia, Didier Dubois, Rudolf Kruse, and Hans-Joachim Lenz, CISM Series, Springer: New York. See also paper 22 at www.probabilityandfinance.com.

[8] Shafer, G. & Vovk, V. (2008). 'A Tutorial on Conformal Prediction'. *Journal of Machine Learning Research*, 9:371-421.

[9] Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic Learning in a Random World*. Springer: New York.

[10] Vovk,V., Nouretdinov, I. & Gammerman, A. On-line Predictive Linear Regression. To appear in *Annals of Statistics*.