The Bactra Review: Occasional and eclectic book reviews by Cosma Shalizi **86**

The Nature of Statistical Learning Theory

by V. N. Vapnik

Berlin: Springer-Verlag, 1995

A Useful Biased Estimator

Vapnik is one of the Big Names in machine learning and statistical inference; this is his statement of ``what is important,'' how to do it, and who figured out how to do it. His views on all these matters are decided, at least a little idiosyncratic, and worth attending to.

The general setting of the problem of statistical learning, according to Vapnik, is as follows. We want to estimate some functional which depends on an unknown distribution over a probability space X --- it could be a ``concept'' in the machine-learning sense, regression coefficients, moments of the distribution, Shannon entropy, etc.; even the distribution itself. We have a class of admissible distributions, called hypotheses, and a ``loss functional,'' an integral over X which tells us, for each hypothesis, how upset we should be when we guess wrong; this implicitly depends on the true distribution. Clearly we want the best hypothesis, the one which minimizes the loss functional --- but to explicitly calculate that we'd need to know the true distribution. Vapnik assumes that we have access to a sequence of independent random variables, all drawn from the (stationary) true distribution. What then are we to do?

Vapnik's answer takes two parts. The first has to do with ``empirical risk minimization'': approximate the true, but unknown, loss functional, which is an integral over the whole space X, with a sum over the observed data-points, and go with the hypothesis that minimizes this ``empirical risk''; call this, though Vapnik doesn't, the ERM hypothesis. It's possible that the ERM hypothesis will do badly in the future, because we blundered into unrepresentative data, but we can show necessary and sufficient conditions for the loss of the ERM hypothesis to converge in probability to the loss of the best hypothesis. Moreover, we can prove that under certain very broad conditions, that if we just collect enough data-points, then the loss of the ERM hypothesis is, with high probability, within a certain additive distance (``confidence interval'' --- Vapnik's scare-quotes) of the loss of the best hypothesis. These conditions involve the Vapnik-Chervonenkis dimension, and a related quantity called the Vapnik-Chervonenkis entropy. Very remarkably, we can even calculate how much data we need to get a given approximation, at a given level of confidence, *regardless* of what the true distribution is, i.e. we can calculate distribution-independent bounds. (They do, however, depend on the nature of the integrands in the loss functional.)

As Vapnik points out, these results about convergence, approximation, etc. are in essence extensions of the Law of Large Numbers to spaces of functions. As such (though he does not point this out), the assumption that successive data-points are independent and identically distributed is key to the whole exercise. He doesn't talk about what to do when this assumption fails.

The second part of Vapnik's procedure is an elaboration of the first: For a given amount of data, we

pick the hypothesis which minimizes the sum of the empirical risk and the ``confidence interval'' about it. He calls this ``structural risk minimization,'' though to be honest I couldn't tell you what structure he has in mind. More popular principles of inference --- maximum likelihood, Bayesianism, and minimum description length --- are all weighed in the balance against structural risk minimization and found more or less wanting.

Vapnik's view of the history of the field is considerably more idiosyncratic than most of his opinions: in epitome, it is that everything important was done by himself and Chervonenkis in the late 1960s and early 1970s, and that everyone else, American computer scientists especially, are a bunch of wankers. Indeed, this is a very Russian book in several senses. I don't just mean that it clearly wasn't written (or edited) by somebody fluent in English --- the missing articles, dropped copulas, and mangled verb-tenses are annoying but not so bad as to conceal Vapnik's meaning. More important, and more characteristically ``Russian,'' is the emphasis on mathematical abstraction, logical rigor and formal elaboration, all for their own sweet sakes. (Detailed proofs are, however, left to his papers.) Vapnik opposes the idea that ``complex theories don't work, simple algorithms do,'' which is fair enough, but he seems almost hurt that simple algorithms ever work, that something as pragmatic and unanalytical as a neural network can not just work, but sometimes even outperforms machines based on his own principles. There are a number of other oddities here, like an identification of Karl Popper's notion of ``unfalsifiable'' with classes of functions with infinite VC dimension, and some talk about Hegel I didn't even try to understand.

I think Vapnik suffers from a certain degree of self-misunderstanding in calling this a summary of learning theory, since many issues which would loom large in a general theory of learning --- computational tractability, chosing the class of admissible hypotheses, representations of hypotheses and how the means of representation may change, etc. --- are just left out. Instead this is a excellent overview of a certain sort of *statistical inference*, a generalization of the classical theory of estimation. In the hands of a master like Vapnik, this covers such a surprisingly large territory that it's almost no wonder he imagines it extends over the entire field. That said, there *is* a lot here for those interested in even the most general and empirical aspects of learning and inference, though they'll need a strong grasp of mathematical statistics.

Errata: As I mentioned, there are many small grammatical errors, and also errors in citations, e.g. Valiant's 1984 *Communications of the ACM* paper is about a ``Theory of the Learnable,'' not ``Learnability.'' These are unchanged in the second, and supposedly corrected, 1998 printing. I'd provide a complete list, but I neglected to note them as I read.

---

xv + 188 pp., numerous black and white graphs, bibliography, index of subjects
Probability and Statistics
Currently in print as a hardback, US$64.95, ISBN 0387945598. LoC Q325.7 V37

---

10 May 1999
Thanks to Jim Crutchfield for lending me his copy