

Book Reviews

Neural Network Learning: Theoretical Foundations

A Review

John Shawe-Taylor

The scientific method aims to derive mathematical models that help us to understand and exploit phenomena, whether they be natural or human made. Machine learning, and more particularly learning with neural networks, can be viewed as just such a phenomenon. Frequently remarkable performance is obtained by training networks to perform relatively complex AI tasks. Despite this success, most practitioners would readily admit that they are far from fully understanding why and, more importantly, when the techniques can be expected to be effective. The need for a fuller theoretical analysis and understanding of their performance has been a major research objective for the last decade. *Neural Network Learning: Theoretical Foundations* reports on important developments that have been made toward this goal within the computational learning theory framework.

Results from computational learning theory typically make fewer assumptions and, therefore, stronger statements than, for example, a Bayesian analysis. This generality can be both a strength and a weakness. Its strength is in the general applicability of the results. However, its weakness follows because a more general result must be more pessimistic to still hold true in the worst case. A similar differ-

ence exists between parametric and nonparametric statistical tests. Parametric tests are only valid if the data satisfy certain assumptions. If these assumptions hold, they will, however, typically give more accurate results. The analysis of statistical learning theory has very much the flavor of a nonparametric statistical test. Almost no

***Neural Network Learning:
Theoretical Foundations,
Martin Anthony and Peter L. Bartlett, Cambridge University Press, Cambridge, U.K., 1999, 389 pp., ISBN 0-521-57353-X, \$59.95 (hardcover).***

assumptions are made about the distribution generating the data. In addition, its bounds hold with high probability in the same way that significance in a statistical test indicates the probability that the data have misled you into accepting a particular hypothesis. For this reason, computational learning theory results are often referred to as probably (that is, with high proba-

Book Reviewers Wanted!

If you are interested in reviewing books for *AI Magazine*, please visit www.cis.ohio-state.edu/~chandra/books-needing-reviewers.htm, and contact chandra@cis.ohio-state.edu.

bility or significance) approximately correct (that is the generalization error is low), or *pac*. The weakness of *pac*, therefore, is that its results must hold true even in worst-case distributions.

There is, however, a new twist to this story in that the more recent *pac*-style results are able to take account of observed attributes of the function that has been chosen by the learner, for example, its margin on the training set. Such attributes measure how beneficial the particular distribution is and feed directly into the bound on the generalization, hence helping to motivate learning strategies that attempt to minimize the particular bound, for example, by maximizing the margin. For this reason, the new style of analysis is often referred to as *data dependent*. Bartlett and Anthony have been two of the researchers driving these developments and, hence, are particularly well placed to produce a book, one of whose main goals is to show the reader how these new results affect neural network learning.

The first part of the book looks at classification using binary-output neural networks. The approach presented is the "classical" (non-data-dependent) *pac* learning analysis of binary classifiers based on the Vapnik-Chervonenkis dimension and associated growth function. A thorough coverage is given of these results, including proofs of all the main theorems. An alternative proof of Sauer's lemma owed to Steele (1978) is given, and a detailed description is included of the crucial symmetrization lemma that forms the core of the Vapnik-Chervonenkis theorem. It is clear by this point that the book aims to give a comprehensive account not only of the results but also of their detailed proofs. This thorough-

ness extends to the lower bounds on the sample complexity in terms of the Vapnik-Chervonenkis dimension and even to the bound on the Vapnik-Chervonenkis dimension of sigmoidal neural networks following Karpinski and Macintyre (1997). Without doubt, the presentation is the fullest exposition of this collection of results in a single text.

Part two turns to using real-valued functions for classification. The book aims to break new ground here by introducing the recent data-dependent results relating the generalization of such a classifier to its margin on the training set or, more generally, the margin together with the number of points failing to meet this margin. Viewing a classifier at a certain scale of margin effectively lowers the complexity of the function class for the particular learning problem. This approach, however, does not correspond to choosing a simple subclass of functions because there is no restriction on which function from the full class can be chosen. It is as though having chosen the solution, we view the class through a margin-scaled filter. Hence, we only need to find functions that can approximate the behavior of the class on the training set to a scale proportional to the margin. Such a set of exemplar functions is known as a *cover*. The log of the size of the cover plays the role of the Vapnik-Chervonenkis dimension in bounding the generalization, and for large margins, this value can be significantly smaller than the Vapnik-Chervonenkis dimension. Furthermore, the Vapnik-Chervonenkis dimension can be replaced by a scale-sensitive version known as the *fat-shattering*, or P_γ dimension, which can be used to bound the size of the covers in a manner analogous to that in which Sauer's lemma bounds the growth function in terms of the Vapnik-Chervonenkis dimension.

Once again, the book gives us the full treatment of these results, including the bounding of the size of the covers of multilayer neural networks with weights bounded in a manner reminiscent of weight decay. Thus, the authors can provide a theoretical justification for the use of weight decay in neural network training to create a

large separation between the positive and negative training examples.

The third part of the book deals with learning real-valued functions, a task often referred to as *regression*. It is possible to treat regression with the tools developed for classification if we are happy to bound the probabilities that a regressor makes an error greater than some threshold on a randomly drawn test point, an approach referred to by the authors as *approximate interpolation*. The authors initially prefer to consider the more usual measure of the expected quadratic loss, that is, $E(f(x) - y)^2$ as the generalization error of the function f , where the expectation is over the distribution generating the data pairs (x, y) . This analysis requires the use of covers over different metrics, although the form of the results is reminiscent of those for classification using real-valued functions, with the fat-shattering dimension playing a central role. Once again, results for neural networks suggest the use of weight decay as a strategy for controlling the capacity of the network without having to reduce its size or the number of weights. There are short sections dealing with convex classes for which tighter bounds can be derived with general loss functions, multiple output networks, and the approximate interpolation approach mentioned earlier.

The final part of the book deals with the question of algorithmics. It was Valiant in his seminal paper who first placed algorithmic efficiency as a core requirement of learning theory. The first three parts of the book ignored this question, preferring to concentrate on estimating how much data are required to obtain good generalization with high probability under the assumption that we can make the best use of these data. In practice, we must be able to find a hypothesis or, in the case of neural networks, a weight setting that optimizes the derived criteria, for example, maximizes the margin but keeps the size of the weights controlled.

Unfortunately, the so-called loading problem for neural networks is hard under the normal complexity assumption that $RP \neq NP$. These results are described before going on to consider the single neuron or perceptron. Even

in this case, the problem of minimizing the number of training errors is hard. Only by restricting consideration to fixed fan-in perceptrons can efficient learning algorithms be described; although even in this case, they have the flavor of enumerating all possible dichotomies that can be realized, an approach that is unlikely to be practical for realistically sized problems.

The book, however, is able to offer more positive results for convex combination constructive algorithms when learning real-valued functions because it can be shown that a good approximation to the global optimum can be found, provided each iteration uses a close to optimal component. This result has a close relationship to the results for sample complexity of convex combinations and lead into a discussion of boosting algorithms for classification, another example of an efficient procedure that is guaranteed to give small training error and exploits the margin ideas to frequently give good generalization error.

Anthony and Bartlett have given us the most thorough treatment of the statistical analysis of neural network learning available to date. They have presented a complete picture of how the proofs are derived right down to an appendix listing the background results that are used in their derivations. The book is therefore an invaluable reference for the learning theorist, at the same time providing the first full treatment of the data-dependent analysis that has brought learning theory significantly closer to the practitioner. Although it is too soon to expect the actual results to provide realistic bounds on the generalization of particular classifiers, the form of the results is already able to motivate algorithmic strategies that frequently do improve generalization performance.

John Shawe-Taylor obtained a Ph.D. in mathematics at Royal Holloway, University of London, in 1986. He subsequently completed an MSc in the foundations of advanced information technology at Imperial College. He was promoted to professor of computing science in 1996. He has published over 100 research papers.