Lecture Notes in Statistics **118**
Edited by P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg,
I. Olkin, N. Wermuth, S. Zeger

Springer Science+Business Media, LLC

Radford M. Neal

Bayesian Learning for Neural Networks

Springer

Radford M. Neal
Department of Statistics and
Department of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 1A4

ISBN 978-0-387-94724-2 ISBN 978-1-4612-0745-0 (eBook)
DOI 10.1007/978-1-4612-0745-0

CIP data available.
Printed on acid-free paper.

Camera ready copy provided by the author.

9 8 7 6 5 4 3

Preface

This book, a revision of my Ph.D. thesis,[1] explores the Bayesian approach to learning flexible statistical models based on what are known as "neural networks". These models are now commonly used for many applications, but understanding why they (sometimes) work well, and how they can best be employed is still a matter for research. My aim in the work reported here is two-fold — to show that a Bayesian approach to learning these models can yield theoretical insights, and to show also that it can be useful in practice. The strategy for dealing with complexity that I advocate here for neural network models can also be applied to other complex Bayesian models, as can the computational methods that I employ.

In Chapter 1, I introduce the Bayesian framework for learning, the neural network models that will be examined, and the Markov chain Monte Carlo methods on which the implementation is based. This presentation presupposes only that the reader possesses a basic statistical background.

Chapter 1 also introduces the major themes of this book, which involve two fundamental characteristics of Bayesian learning. First, Bayesian learning starts with a prior probability distribution for model parameters, which is supposed to capture our beliefs about the problem derived from background knowledge. Second, Bayesian predictions are not based on a single estimate for the model parameters, but rather are found by integrating the

---

[1] *Bayesian Learning for Neural Networks*, Department of Computer Science, University of Toronto, 1995.

model's predictions with respect to the posterior parameter distribution that we obtain when we update the prior to take account of the data. For neural network models, both these aspects present difficulties — the prior over network parameters has no obvious relation to any prior knowledge we are likely to have, and integration over the posterior distribution is computationally very demanding.

I address the first of these problems in Chapter 2, by defining classes of prior distributions for network parameters that reach sensible limits as the size of the network goes to infinity. In this limit, the properties of these priors can be elucidated. Some priors converge to Gaussian processes, in which functions computed by the network may be smooth, Brownian, or fractionally Brownian. Other priors converge to non-Gaussian stable processes. Interesting effects are obtained by combining priors of both sorts in networks with more than one hidden layer. This work shows that within the Bayesian framework there is no theoretical need to limit the complexity of neural network models. Indeed, limiting complexity is likely to conflict with our prior beliefs, and can therefore be justified only to the extent that it is necessary for computational reasons.

The computational problem of integrating over the posterior distribution is addressed in Chapter 3, using Markov chain Monte Carlo methods. I demonstrate that the hybrid Monte Carlo algorithm, originally developed for applications in quantum chromodynamics, is superior to the methods based on simple random walks that are widely used in statistical applications at present. The hybrid Monte Carlo method makes the use of complex Bayesian network models possible in practice, though the computation time required can still be substantial.

In Chapter 4, I use a hybrid Monte Carlo implementation to test the performance of Bayesian neural network models on several synthetic and real data sets. Good results are obtained on small data sets when large networks are used in conjunction with priors designed to reach limits as network size increases, confirming that with Bayesian learning one need not restrict the complexity of the network based on the size of the data set. A Bayesian approach is also found to be effective in automatically determining the relevance of inputs.

Finally, in Chapter 5, I draw some conclusions from this work, and briefly discuss related work by myself and others since the completion of the original thesis.

Readers interested in pursuing research in this area may obtain free software implementing the methods, as described in Appendix B. One should note, however, that this software is not intended for use in routine data analysis. The software is also designed only for use on Unix systems.

Of the many people who have contributed to this work, I would like first of all to thank my thesis advisor, Geoffrey Hinton. His enthusiasm for understanding learning, his openness to new ideas, and his ability to provide insightful criticism have made working with him a joy. I am also fortunate to have been part of the research group he has led, and of the wider AI group at the University of Toronto. I would particularly like to thank fellow students Richard Mann, Carl Rasmussen, and Chris Williams for their helpful comments on this work and its precursors. My thanks also go to the present and former members of my Ph.D. committee, Mike Evans, Scott Graham, Rudy Mathon, Demetri Terzopoulos, and Rob Tibshirani.

I am especially pleased to thank David MacKay, whose work on Bayesian learning and its application to neural network models has been an inspiration to me. He has also contributed much to this work through many conversations and e-mail exchanges, which have ranged from the philosophy of Bayesian inference to detailed comments on presentation. I have benefited from discussions with other researchers as well, in particular, Wray Buntine, Brian Ripley, Hans Henrik Thodberg, and David Wolpert.

This work was funded by the Natural Sciences and Engineering Research Council of Canada and by the Information Technology Research Centre. For part of my studies, I was supported by an Ontario Government Scholarship.

Contents

List of Figures