Neural Networks and Machine Learning

July - December 1997

Report from the Organisers: Christopher M. Bishop (Microsoft/Edinburgh), David Haussler (UCSC), Geoffrey E. Hinton (Toronto), Mahesan Niranjan (Cambridge), Leslie G. Valiant (Harvard)

- **Introduction**
- **Organisation**
- **Workshops**
- **Achievements**
- **Acknowledgements**

Introduction

The first mathematical models of artificial neural networks were proposed in 1943 by McCulloch and Pitts, who demonstrated that networks of simple threshold units were capable of universal computation. By the 1960's, neural networks had become the focus of extensive research, much of which was concerned with the abilities of adaptive networks to learn from data sets. This did not, however, lead to large-scale applications, largely as a consequence of fundamental limitations of the algorithms considered at that time. Following many years of relative inactivity, the field experienced rapid growth during the 1980s and 1990s, stimulated by the development of new algorithms which overcame the limitations of earlier approaches, and by the widespread availability of fast computers. During this period the close links between neural network models and the concepts of conventional statistical pattern recognition were clarified, leading to a stronger theoretical foundation for neural network algorithms, as well as to more effective practical exploitation. Theoretical insight into neural network models has also come from computational learning theory, approximation theory, dynamical systems theory and information geometry. These disparate viewpoints often provide complementary insights.

The mathematical foundations of neural networks are currently being studied by several different communities of researchers including computer scientists, statisticians and physicists. One of the aims of the programme *Neural Networks and Machine Learning* at the Isaac Newton Institute was to promote greater interdisciplinary collaboration between these different groups. One aspect in particular, namely the interaction between researchers in mainstream neural computing and in the field of probabilistic graphical models, was especially successful.

Organisation

The overall planning of the programme was undertaken by Christopher Bishop, with the other organisers (as well as some participants) being involved in the arrangement of specific workshops, conferences and seminar series.

The programme was the largest international event of its kind to have been held in the field. Many of world's leading researchers in neural computing participated in the programme, for periods ranging from one or two weeks up to six months, and numerous younger scientists benefited from workshops and tutorials. An overview of the workshops is given below.

Regular seminars, organised by Mahesan Niranjan, were run during the weeks between workshops. Several of these comprised tutorials on specific subjects aimed at non-specialists and these proved to be particularly useful in allowing participants to learn about new topics. It was found that about two or three such seminars per week was an appropriate number in maintaining a sense of ongoing activity in the programme while allowing participants adequate time to pursue their own research or to engage in collaborative work. There was an excellent atmosphere of lively interaction during the programme, and we are aware of a significant number of new collaborations which have emerged.

Two of the seminars in the Newton Institute Seminar Series were presented by workshop participants: *Neural Networks: A Probabilistic Perspective* (Christopher Bishop) and *Statistical Genome Analysis using Hidden Markov Models* (David Haussler).

Social events were organised throughout the programme, and these proved to be invaluable in helping participants to get acquainted. The personal reports produced by participants were generally highly positive. In particular, many of the participants commented on the unique and highly stimulating research environment offered by the Newton Institute, and on the tremendous support provided by the Institute staff.

Workshops

NATO Advanced Study Institute

The first workshop of the programme was a two-week NATO Advanced Study Institute on "Generalization in Neural Networks and Machine Learning" which took place from 4 to 15 August (Director: Christopher Bishop; co-organisers: Joachim Buhmann, Geoffrey Hinton and Michael Jordan). This was heavily over-subscribed and attendance was limited to around 110 by the capacity of lecture room 1. Several complementary perspectives on generalization were covered in this workshop, including both Bayesian and frequentist viewpoints. By reviewing many of the key theoretical concepts underpinning the field of neural computing, the NATO ASI provided an excellent start to the overall programme, as well as giving younger scientists the chance to hear tutorial talks by leading researchers in the field. The scientific and social aspects of the programme were both very successful, and a proceedings volume will be published in the NATO series by Springer towards the end of 1998.

Pulsed Neural Networks

A two-day workshop on *Pulsed Neural Networks* was held on 26 and 27 August, and attracted around 60 people (double the number expected). The majority of artificial neural network models are based on the propagation of continuous variables from one processing unit to the next. In recent years, data from neurobiological experiments has made it increasingly clear that biological neural networks, which communicate through pulses, use the timing of these pulses to transmit information and to perform computation. This realization has stimulated a significant growth of research activity in the area of pulsed neural networks, including theoretical analysis as well as the development of new computational paradigms. As a result of this workshop, it was decided to write a multi-authored book involving twelve invited contributions from workshop speakers (edited by Wolfgang Maass and Christopher Bishop) which will be published by MIT Press in November 1998.

EC Summer School

An EC Summer School on *Probabilistic Graphical Models* was held from 1 to 5 September (organised by Christopher Bishop and Joe Whittaker). This workshop was outstandingly successful and generated considerable enthusiasm from the lecturers and participants. It was run as a mixture of 90 minute tutorials in the mornings and 30 minute advanced research talks in the afternoons and ensured that both lecturers and students could derive maximum benefit. A key feature of this workshop was the interaction between neural computing and graphical model researchers, and is discussed further below.

Analysis of DNA Sequences

A workshop on *Statistical Analysis of DNA and Protein Sequences* (organised by David Haussler and Richard Durbin) took place from 20 to 24 October. This was highly topical and attracted a strong participation. This workshop provided something of a precursor to the 1998 programme on *Biomolecular Function and Evolution in the Context of the Genome Project*.

Bayesian Methods

The final programme conference was on "Bayesian Methods" and was held from 15 to 19 December (organised by Christopher Bishop and Mahesan Niranjan). Again this was very well attended and formed an excellent conclusion to the six month programme. It also provided an opportunity for some of the programme participants to present results of research conducted while resident at the Newton Institute.

Themed weeks

In addition to the main conferences, we also ran four `themed weeks', which tended to be less formal, and hence more interactive, than the larger conferences. The typical number of participants was around 30, and the seminars were generally held in lecture room 2.

The first themed week was on the topic of *Learning in Computer Vision*, and was arranged by Andrew Blake, David Mumford and Alan Yuille. It provided a direct link with the earlier Newton Institute programme on Computer Vision, and many of the participants in the themed week were also participants in that programme.

A week on *Applications of Neural Computing* was run from 3 to 7 November, organised by Lionel Tarassenko and Stephen Roberts. The focus was on the interplay between underlying theory and practical implementations. Presentations covered many of the most exciting and successful practical applications of neural networks.

Finally, the topic of *On-line Learning* was explored in depth during the week of 17 to 21 November, organised by David Saad. This themed week has resulted in a multi-authored book (edited by David Saad) based on the research presented at the Institute.

Achievements

Amongst the many successes of the programme, one of the most notable was the stimulation of new interdisciplinary collaboration between the neural computing and the probabilistic graphical model communities. This was reflected in the highly successful EC workshop, and was promoted by the long-term participation of several key graphical model researchers. The Newton Institute programme represented the first

occasion on which these two communities have interacted so closely over an extended period. A great deal of positive feedback was received from participants both during the EC

workshop and afterwards by email, and the workshop was described by several people as a landmark event.

Thomas Richardson from the University of Washington was elected as Rosenbaum fellow. Amongst the research carried out during his stay at the Institute, he developed a new type of graphical model designed to represent the Markov structure that a latent variable model induces over its observed margin. He showed that Gaussian mixed ancestral graphs are in fact curved exponential family models, and hence such graphs describe a manifold in parameter space, whereas a typical latent variable model does not. As a consequence the Bayesian information criterion (BIC) is consistent for mixed ancestral graph selection, i.e. in the asymptotic limit BIC will assign the highest score to the true model.

Christopher Bishop (Microsoft/Edinburgh), Brendan Frey (Illinois) and Neil Lawrence (Aston) have explored novel approximating distributions for variational inference and learning in densely connected Bayesian networks. Exact inference is intractable in such models, and standard mean field theory addresses this problem by assuming complete factorization. They showed that a much richer approximating distribution, involving a Markov chain framework, also leads to a tractable algorithm.

Joe Whittaker (Lancaster) and Alberto Roverato (Lancaster) developed a new importance sampling approach for the evaluation of normalizing constants in non-decomposable Gaussian graphical models. The sampler is based on the asymptotic distributions of either the maximum likelihood estimates or of the posterior distribution in a conjugate analysis. Two problems had to be addressed. First, the support of the required distribution is that of a positive definite matrix constrained by a particular structure of zeros. (Sampling from standard matrix distributions such as the Wishart would give zero support to the required distribution.) Second, careful numerical implementation is required to ensure that the shape of the required distribution is accurately tracked by the sampler in spaces of high dimensionality.

Milan Studeny (Prague) developed and presented a new, direct separation criterion for chain graphs, and showed that it is equivalent to a classical moralization criterion. A chain graph is a probabilistic graphical model admitting both directed and undirected edges, with (partially) directed cycles forbidden. Using the new criterion he then proved that for every chain graph there exists a strictly positive definite probability distribution that embodies exactly the independency statements displayed by the graph, thereby justifying the use of chain graphs as a tool for the description of conditional independence structures.

Tommi Jaakkola (UCSC) and David Haussler (UCSC), motivated by problems in biosequence analysis, have developed novel kernel-based methods which combine discriminative and generative approaches. Biosequence analysis relies to a large degree on the assessment of similarity between DNA or protein sequences. It is therefore important that the statistical techniques employed for the analysis make this similarity measure explicit. Several (discriminative) methods, such as support vector machines and Gaussian process classifiers, provide means for explicitly characterizing the similarity metric. While certainly powerful, these methods nevertheless lack the ability to deal well with the inherent variability of biosequences for which generative statistical models in turn are more appropriate. Ideally, these complementary approaches should be combined. This research has indeed led to the development of a general framework for combining generative models with discriminative (kernel) methods, motivated by considerations from information geometry, thus establishing a new class of statistical tools.

Mahesan Niranjan continued his work on Bayesian methods applied to signal processing problems, particularly the problem of speech enhancement. Using Bayesian methods he

showed how to estimate the noise statistics from corrupted data without the need to segment the speech from the background noise. This leads to an algorithm with which it is possible to enhance speech corrupted by white noise, and the approach permits a natural extension to deal with coloured noise.

David Haussler (UCSC) and Manfred Opper (Aston) considered the problem of sequential prediction under log loss, in a scenario in which a player tries to minimize his loss relative to the loss of the (with hindsight) best distribution from a target class for the worst sequence of data. They obtained bounds on the minimax regret in terms of the metric entropies of the target class with respect to suitable distances between distributions. In addition, they showed that such a worst case scenario leads to prediction strategies which also work well in a less pessimistic, average case situation provided the model class is not too complex.

Shun-ichi Amari (Tokyo) visited the Institute for three weeks and presented some of his recent work on the information geometry view of learning in neural networks and other probabilistic models, including a unified perspective on the various techniques for independent component analysis. One of the developments to have emerged from this work is the use of the natural Riemannian gradient for on-line learning. It is anticipated that this algorithm is not only asymptotically efficient but may eliminate the so-called plateaus associated with conventional gradient descent learning which give rise to extremely slow convergence. During Amari's visit, Magnus Rattrey (Aston) and David Saad (Aston), in collaboration with Amari, developed an analysis of the performance of natural gradient descent within the statistical mechanics framework, and were able to go beyond the asymptotic regime. The Riemannian gradient has also been applied to the problem of independent component analysis in the case where the number of source signals is unknown. Usually the number of observations is larger than that of the sources but they are noise contaminated. In this case the parameter space forms a Stiffel manifold, and an explicit formula has been obtained for the natural gradient.

Acknowledgements