Seconder of the vote of thanks and contribution to the Discussion of 'the Discussion Meeting on Probabilistic and statistical aspects of machine learning'

Christopher Nemeth

September 2023

I congratulate the authors of these two papers for their insightful and significant contributions to addressing the statistical aspects of machine learning. In this contribution to these discussion papers, I will make a short comment which covers both papers and then provide some separate thoughts on each paper.

1 Statistical Aspects of Machine Learning

These two papers are quite different in their focus, however, a common thread between them is the use of neural networks, and in particular deep neural networks, to augment part of the modelling process. In the case of Li et al. (2022), neural networks are used to convert the changepoint problem into a supervised learning problem, and in the case of Benton et al. (2022) neural networks are used to approximate the intractable score function.

A better understanding of the statistical properties of neural networks is an ongoing area of research (Anthony et al., 1999; Bartlett et al., 2019), but their application has become widespread within the artificial intelligence (AI) and machine learning communities. Within the statistics community, we can look back 30 years to the discussion paper of Ripley (1994) to see how neural networks can be used to solve classification problems. Interestingly, although the Ripley (1994) and Li et al. (2022) papers are very different in their focus, they both utilise neural networks to solve a classification problem and derive similar theoretical results regarding neural network complexity in terms of VC-dimension bounds.

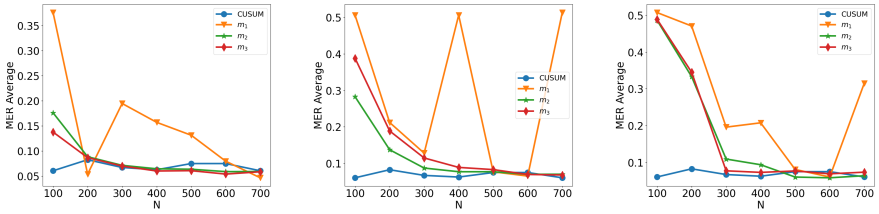

Figure 1: Scenario 1: $n = 100, N \in \{100, 200, \ldots, 700\}, \rho = 0$. ReLU (left), SeLU (middle) and Sigmoid (right) activation functions.

2 Paper 1: Automatic Change-point Detection in Time Series via Deep Learning by Li et al.

The authors comment in the conclusion to their paper that for applied statisticians trying to model a changepoint problem: "*...they need to understand what type of change is sought, be able to characterise it mathematically, find a satisfactory stochastic model for the data, formulate the appropriate statistic, and fine-tune its parameters.*" However, if instead of trying to model the data-generating process a neural network is used, then does this not lead to the same challenges? What type of neural network should be used? How deep should the network be? As an illustration of this point, let us consider the example from Section 5 of the paper, with $\rho = 0$ under Scenario 1. Figure 1 presented here is similar to Figure 2 in Li et al. (2022) where I have changed the activation function from ReLU (left) to SeLU (middle) and sigmoid (right). Under these different activation functions, we can see that for sufficiently large N, the neural network-based approaches are superior to the CUSUM statistics, but for small N the average misclassification error rate can be quite different depending on the choice of the activation function. What guidelines are available for practitioners to ensure that they use the best neural network architecture? Is it easier to choose an appropriate neural net than it is to choose a statistical model that directly models the time series data?

The authors focus their empirical presentation on the statistical improvements of their neural network-based approach when compared against the CUSUM statistic. However, there is no presentation of the difference in computational cost between the CUSUM and neural network approaches. Would it be more reasonable to report the misclassification error rate scaled by computational time? This would be interesting to consider because if it takes twice as long, or perhaps longer, to fit the neural network compared to the CUSUM test, then what percentage of improvement should we expect to see from the neural network as a result of the increased computational complexity?

3 Paper 2: From Denoising Diffusions to Denoising Markov Models by Benton et al.

The general denoising framework proposed in this paper is an important contribution that allows the class of diffusion generative models, which are rapidly growing in popularity, to be applied to non-Euclidean spaces. The examples in the paper are directed towards sampling from non-standard spaces, however, the first example, which considers an approximate Bayesian inference model, is on $\mathbb{R}^d$ and provides a nice illustration of how these diffusion models are applied on simpler problems. What is of particular interest in Section 6.1 is that the results presented are not significantly better than many existing approximate Bayesian computation (ABC) algorithms. By some standards this is quite a simple example as there are only four parameters to be learnt, and yet the neural network (a multilayer perceptron) used by the authors to approximate the score function has 1.9 million parameters (see Appendix J.1). It does seem somewhat paradoxical that in order to approximate a 4-dimensional distribution it is necessary to estimate 1.9 million parameters in a neural network. Furthermore, the observed dataset is of size 250, which leads to interesting questions around how feasible it is to learn a large number of parameters from a neural network with small datasets. Is it possible to know what type of neural network should be used for a particular generative problem, e.g. images, text, etc.? Or how large the network needs to be in order to achieve high-levels of statistical accuracy?

In conclusion, these two papers provide stimulating contributions to the field of statistical machine learning and open up many interesting avenues of future research in these areas, it is a pleasure to second the vote of thanks.

References

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. (1999). *Neural network learning: Theoretical foundations*, volume 9. Cambridge University Press Cambridge.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301.

Benton, J., Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. (2022). From denoising diffusions to denoising markov models. *arXiv preprint arXiv:2211.03595*.

Li, J., Fearnhead, P., Fryzlewicz, P., and Wang, T. (2022). Automatic change-point detection in time series via deep learning. *arXiv preprint arXiv:2211.03860*.

Ripley, B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):409–437.