Discussion of two papers on Probabilistic and Statistical Aspects of Machine Learning

RSS 2023 Conference, Harrogate, UK (6th September, 2023)

Darren Wilkinson (Durham University)

darrenjw.github.io

Introduction

Overview

- Computational statistics and machine learning are closely related, and in many cases are not even clearly distinct
- Many opportunities for cross-fertilisation of ideas from the two fields
- Ideas and methods from probability and statistics have been very important in the development of modern ML models and algorithms (especially deep generative models)
- Increasingly, ideas and methods from ML are finding their way into statistics
- Both fields can benefit from greater interaction, and the two papers being discussed today highlight a few of the ways that this can happen

Paper 1: Automatic Change-Point Detection in Time Series via Deep Learning (Li et al)

Automatic Change-Point Detection in Time Series via Deep Learning

- Offline change point detection (main focus on a single change point)
- Using (a lot of) labelled training data — many examples like the problem of interest, where the true change point has been manually labelled
- Interest is in the automatic generation of new offline detection methods using neural networks
- Providing statistical guarantees of method performance
- Theory is developed for a class of MLPs that directly generalise existing CUSUM-based methods
- Little previous work using historic labelled data to develop offline change-point methods — why?

Observations and questions (1)

- The theory applies to a basic MLP with ReLU activation — standard model, amenable to statistical analysis, and nicely generalises existing methods
- Examples all use MLPs of constant layer width — rarely seen in practice
- Practical advice on choosing network depth and layer widths sensibly and safely?
- How to interpret the width condition: $m_r m_{r+1} = \mathcal{O}(n \log n)$?
- Need a lot of training data — for a wide single layer MLP, need $N >> n^2 \log n$
- For more realistic architecture, more like $N >> n \log^2 n$?
- Min-max scaling might be bad in the presence of heavy-tailed noise?

Observations and questions (2)

- For the activity data application, a more SOTA neural network architecture is adopted (CNN layers, residual blocks, skip connections, etc.)
- This makes perfect sense, but the theory doesn't directly apply
- What hope is there of extending theory to very general model classes?
- What practical advice can be given?
- In the absence of labelled data, but the presence of a full data generating process (including change-points), a simulator can be used to train the network
- This is then exactly the framework of *simulation-based inference* (SBI)
- Standard trade-off: strong modelling assumptions reduce the need for large training data sets

Paper 2: From Denoising Diffusions to Denoising Markov Models (Benton et al)

From Denoising Diffusions to Denoising Markov Models

- Denoising diffusions (and related models) are deep generative models for simulating (conditional) samples from a data distribution
- Huge training data sets are required, but again, these can be replaced by a data-generating process in the context of *simulation-based inference* (eg. first experiment in the paper)
- Typical applications are to very high dimensional data (eg. images), but can also be used for sampling Bayesian posterior distributions (eg. first experiment)
- Expensive both to train and sample (even relative to other deep generative models), but are SOTA for certain problems

Denoising Markov models

- The paper provides a unifying mathematical framework for a broad class of models of the denoising diffusion form (with fairly arbitrary state spaces)
- The emphasis is on the formulation in continuous time (potentially beneficial for both practical and theoretical reasons), but the connection with discrete time formulations is clearly articulated
- Conditional simulation (including Bayesian posterior sampling) is briefly discussed
- The approach is to work with continuous time Markov processes on a general state space, and to formulate the (de)noising process in terms of the infinitesimal generator of the Markov process
- The resulting optimisation targets are shown to generalise several different (sometimes *ad hoc*) special cases that have appeared in the literature for particular state spaces

Observations and questions (1)

- The emphasis is on a unifying framework, but no discussion in the (main) paper of how to actually generate samples (simulate realisations from the reversed process)
- This is very important in practice, but also very dependent on the state space — are there any general observations that can be made?
- The examples described in the supplementary materials use approximate first order methods based on a fine regular time grid — probably not optimal
- One of the attractions of formulating the models in continuous time is the possibility of using higher-order methods with adaptive time steps
- Might adaptive time-stepping reduce the need for time-rescaling?
- For some applications it is convenient to have a deterministic generation mechanism (given an initial noise sample), using some sort of probability flow differential equation — is such an approach covered by the general denoising Markov model framework presented here?

Observations and questions (2)

- Everything depends on using a "good" neural network architecture for the (conditional) denoising process — can anything general be said about how to choose the architecture for a given problem?
- Can anything be said about the kinds of problems for which denoising Markov models work well, and when they don't?
- Why aren't these models more widely used for simulation-based inference? Are DMMs competitive here?
- The g-and-k example wasn't especially impressive (eg. Figure 8 in the supplementary), despite being a fairly standard low-dimensional Bayesian inference problem — could issues be diagnosed without ground truth, and can it be fixed?
- The framework in the paper seems to cover state spaces with mixed discrete and continuous variables — are there good examples of denoising Markov models being used for such problems?

Reflections

Probabilistic and Statistical Aspects of Machine Learning

- These two impressive papers illustrate two quite different aspects of the interaction between statistics and ML
- Caricaturing somewhat, Paper 1 can be seen more in terms of bringing new ideas from ML into statistical modelling, whereas Paper 2 is more about using ideas from probability theory and statistics to advance ML
- From the perspective of academic statistics, we will see increasing use of modern ML methods in statistical methodology papers
- The ML methods are powerful and flexible, but need a lot of training data (and compute)
- Such methods are natural in the context of "big data", but are also useful for "small data" in conjunction with a fully specified data-generation process (simulation-based inference)

Looking forward

- Over the next decade, it is likely to become increasingly difficult to draw a clear line between computational statistics and ML, but this comes with challenges
- Although the methods used by (say) computational Bayesian statisticians and deep generative ML researchers are already barely distinguishable, the language and culture of the two communities remains quite distinct — there is a communication barrier
- Programming languages also illustrate potential issues — Python is the language typically used for ML and associated frameworks such as TensorFlow (used for Paper1), JAX (used for Paper 2), and Torch — yet most academic statisticians currently use R by default (in fact, R and Python are both hopelessly inadequate for scalable statistical computation and ML, but that is another story)

In conclusion

- ML is a fast-moving field closely related to computational statistics
- The opportunities for cross-fertilisation of ideas between statistics and ML is great, and growing
- The two papers presented today are important contributions in their own right, but also serve to highlight some of the potential benefits of narrowing the gap between the two communities
- **It therefore gives me great pleasure to propose the vote of thanks!**