Modeling Streaming Data In the Absence of Sufficiency

Frank Wood
Columbia University, New York, NY 10027, USA
fwood@stat.columbia.edu

Abstract

We interpret results from a study where data was modeled using constant space approximations to the sequence memoizer. The sequence memoizer (SM) is a non-constant-space, Bayesian nonparametric model in which the data are the sufficient statistic in the streaming setting. We review approximations to the probabilistic model underpinning the SM that yield the computational asymptotic complexities necessary for modeling very large (streaming) datasets with fixed computational resource. Results from modeling a benchmark corpus are shown for both the effectively parametric, approximate models and the fully nonparametric SM. We find that the approximations perform nearly as well in terms of predictive likelihood. We argue from this single example that, due to the lack of sufficiency, Bayesian nonparametric models may, in general, not be suitable as models of streaming data, and propose that nonstationary parametric models and estimators for the same inspired by Bayesian nonparametric models may be worth investigating more fully.

1 Introduction

The sequence memoizer [Wood et al., 2009, 2011] (SM) is a Bayesian nonparametric model for sequential stochastic processes that generate discrete observations. Because the SM is a model in which the data is the sufficient statistic, it has space complexity that grows linearly as a function of the number of observations. Since it was first introduced, two complementary modifications to the sequence memoizer have emerged [Bartlett et al., 2010; Gasthaus and Teh, 2011] that, when combined as they were in [Bartlett and Wood, 2011], together result in a constant space approximation to the sequence memoizer.

To review: the SM can be thought of as a hierarchical smoothing prior for multiple simultaneous conditional distribution estimation. "Observations" in the SM are the discrete (countable) symbols generated by some underlying stochastic process and situated in the full sequence or "context" of observations that were already generated. It has been shown that the space complexity of the sequence memoizer is on the order of the number of nodes in a suffix-tree representation of this sequence of observations times the storage required to represent the conditional density estimate at each node. An approximation to the sequence memoizer in which the number of nodes in the suffix-tree remains of asymptotically constant order was introduced in [Bartlett et al., 2010]. Unfortunately, in that work the storage requirement at each node grew as an uncharacterized but not-constant function of the input sequence length. A method for constraining the memory to be of constant order at each node was later introduced in [Gasthaus and Teh, 2011]. Unfortunately, the representation they proposed resulted in the computational cost of inference growing as a super-linear function of the length of the training sequence. Bartlett and Wood [2011] combined both and introduced two more approximations that rendered the computational cost of inference asymptotically linear in the observation sequence length and cost of storage asymptotically constant in the same. We intrepret results from that paper here.

2 Experiments

In order to achieve asymptotically constant storage and linear time cost to generate a single posterior sample, the number L of conditional distribution estimates maintained, the maximum number k of observations held at each node, the length of a buffer T holding observations all must be capped. Optionally the maximum

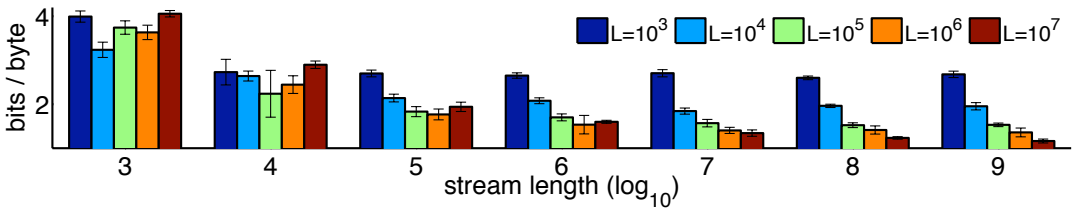

Figure 1: Average ($\pm$ std.) SM predictive performance in terms of average number of bits required to encode each byte of uncompressed input. Here the input stream length and number of distinctly estimated conditional distributions (L) are varied. Observe that average performance monotonically improves as a function of both the input sequence length and of the exactness of the approximation. That performance varies as a function of model complexity for the same size data suggests that the approximate posterior used to compute average model performance may not accurately average over uncertainty about the larger models' parameters.

context length D can also be capped. The resulting approximate SM can be interpreted as a semi-greedy approximation to the full SM in which the number of represented conditional density estimates is constrained to be simultaneously less than both the number of nodes in the suffix tree representation of the "remembered" sequence of observations of length T and the hard constraint on the number of conditional distribution estimates L. It can also be interpreted as a dependent nonparametric model [Bartlett et al., 2010]. The performance of SM-approximations have been explored by modeling subsections of length L sampled with replacement from the complete Wikipedia text content dump [Wikipedia, 2010]. (Figure 1). In these experiment $k = 8,192$, $T = 10^8$, and $D = 16$ while L varied as shown in the figure. The largest value of L has performance indistinguishable from that of the full sequence memoizer for values of stream length up to 10^7. The full sequence memoizer could not be instantiated for the largest datasets so no comparison can be provided.

3 Discussion

Whac-a-Mole^{TM} is a frustrating yet fun game common to American video-game parlors. Game play consists of using a hammer to bash the head of a mole that randomly pops up out of an array of holes in front of the player. Our experience with scaling the sequence memoizer was like a game of whac-a-mole with the nonparametric nature of the SM playing the role of the mole – it kept popping up and ruining our attempts, sometimes in unexpected ways, despite all our bashing. Ultimately we succeeded in producing an asymptotically scalable version of the SM for streaming data, however, to do so we had to abandon essentially all of its nonparametric features. Intuitively the result can be thought of as some kind of estimator for a non-stationary, parametric model; however the exact nature of this model has yet to be characterized. The lessons learned by playing this game are obvious in retrospect and extend to all fundamentally nonparametric models where the data is the sufficient statistic: asymptotic scalability is a problem. For testing and statistical discovery of patterns in finite datasets, perhaps Bayesian nonparametric models have a role to play; however, our experience suggests a no-free-lunch result: it seems that one can not abandon sufficiency and have scalability too.

References

Bartlett, N., Pfau, D., and Wood, F. (2010). Forgetting counts: Constant memory inference for a dependent hierarchical Pitman-Yor process. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 63–70.

Bartlett, N. and Wood, F. (2011). Deplump for streaming data. In *Data Compression Conference*, pages 363–372.

Gasthaus, J. and Teh, Y. W. (2011). Improvements to the sequence memoizer. In *Proceedings of Neural Information Processing Systems*, pages 685–693.

Wikipedia (2010). URL: `http://download.wikimedia.org/enwiki/`.

Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. W. (2009). A stochastic memoizer for sequence data. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1129–1136, Montreal, Canada.

Wood, F., Gasthaus, J., Archambeau, C., James, L., and Teh, Y. (2011). The sequence memoizer. *Communications of the ACM*, 54(2):91–98.