Research Statement

Nicholas Bartlett

Traditional parametric statistical tools and methods are designed to allow inference about a population from a small sample. In parametric models inference is limited to a finite set of parameters and as the amount of data grows, posterior uncertainty about the value of the parameters will typically vanish, rendering the inclusion of additional data irrelevant. However, in nearly all naturally arising data it is understood that a parametric model only approximates the true generative process and thus this convergence of estimators is unsatisfactory. It is our contention that with access to large scale data we should attempt to model complex phenomena by increasing the representational capacity, which usually means complexity, of the models we use.

Sometimes one has access to so much data that parametric models themselves are not necessary. In other words, for some problems one can do strict nonparametric inference; i.e. one can query the data directly. For such nonparametric approaches to have good practical performance the size of the data must be large relative to the complexity of the model. We suggest that there are naturally occurring stochastic processes of sufficient complexity (e.g. natural language generators) that the enormous size of any reasonable model renders even seemingly gigantic data too small for this approach (e.g. non-smoothed n-gram language models). To attack these kinds of problems, we advocate and actively pursue computationally practical ways to estimate and perform inference in Bayesian nonparametric (BNP) models. BNP models are nonparametric in nature, which gives them representational capacity that can be understood to grow as a function of the amount of training data. BNP models are also Bayesian in nature which allows for hierarchical regularization and incremental inference and estimation. For small scale data on the order of millions of observations, BNP natural language models and lossless compressors we have recently been shown to exhibit excellent empirical characteristics [3, 4, 2]. Unfortunately BNP models in general have been saddled with an unfortunate stigma, namely that they are as a class uniformly computationally complex. We suspect that this stigma is at least partially responsible for holding back wide adoption of BNP methods.

Our current path of research aims to harness the power of BNP models with computationally tractable algorithms. We aim to develop linear time (in the length of the data) inference mechanisms that permit constant time prediction and allow for constant space representation. We contend that without these computational constraints BNP algorithms will remain impractical in many interesting applications where the scale of the data is prohibitive. An example of our efforts is the development of a dependent hierarchical Pitman-Yor process model for discrete sequence prediction [1]. As more complex methods are developed and shown to be effective it will become increasingly important to facilitate their deployment on massive scale data.

References

[1] Bartlett, N., Pfau, D., and Wood, F. (2010). Forgetting counts : Constant memory inference for a dependent hierarchical Pitman-Yor process. In *(*to appear) ICML.

[2] Gasthaus, J., Wood, F., and Teh, Y. W. (2010). Lossless compression based on the Sequence Memoizer. In *Data Compression Conference 2010*, pages 337–345.

[3] Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the Association for Computational Linguistics*, pages 985–992.

[4] 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.