Hierarchical Bayesian Models for Artificial Intelligence

PI	Frank Wood, Ph.D.	**Position**	Assistant Professor
Address	Room 1005 SSW, MC 4690 1255 Amsterdam Avenue New York, NY 10027	**University** **Department**	Columbia University Statistics
Phone	212.851.2132	**Fax**	212.851.2164
Website	http://www.stat.columbia.edu/∼fwood		

Abstract

We propose to test the hypothesis that hierarchical Bayesian nonparametric (BNP) models can be used to learn compact world representations suitable for efficient artificial intelligence (AI) search and planning.

Keywords : hierarchical models, Bayesian nonparametrics

Research Areas : machine learning, unsupervised probabilistic modeling

Artificial Intelligence

Artificial intelligence is arguably best posed as a search problem [9]. Given a goal (reproductive success, crossing a street, winning a game of chess or poker), achievement of that goal can be phrased as the problem of searching through a space of possible future worlds to find the action that optimizes the probability or amount of future reward. This setup is very general; almost any planning problem can be posed and solved using this formalism. Unfortunately, for most interesting problems the search space typically gets very big and therefore becomes computationally intractable. Pruning the search space is the only way render the search problem tractable in general. Learning models of the world that explain high-dimensional observations in terms of simpler, unobserved, but sufficiently explanatory latent states is one way to do this. How to do this is arguably the biggest problem that must be solved in order to develop AI. This is the over-arching theme of my work.

Modeling

Probabilistic models are a mathematical framework for describing complex observations as the noise-corrupted output of simpler, often stochastic, latent mechanisms [4]. The learned latent mechanism can be used for efficient AI planning and search.

I focus on learning models that satisfy an additional set of natural life-long learning requirements: models that are representable in constant space, incrementally updatable in constant time, and automatically adaptable to changes in the underlying world dynamics. My focus thus-far has been on models that fall at the bottom of a hierarchy of classical model complexity, but that actually are of infinite complexity; the simplest of which is an infinite Markov model called the sequence memoizer (SM) [10]. A Markov model describes the world as a deterministically transitioning state machine that generates noisy outputs from a known set of states. The classical complexity and expressivity of a Markov model is a function of its number of states. The SM has infinite state cardinality and therefore unbounded complexity, yet satisfies the life-long learning requirements [1, 2, 5]. The SM achieves good performance because of hierarchical sharing of statistical strength between related states.

While Markov models like the SM are often very good models, they would seem to be representationally inefficient as the states in the model must be indexed by all observed features of the world. This state explosion is in conflict with AI model desiderata. The simplest kind of model that can counter this multiplicative state expansion is one that still transitions deterministically from state to state, but where these "states" are abstract, implicit to the model, and no longer in one-to-one correspondence with states of the world. We have defined such a model called the probabilistic deterministic infinite automata, PDIA [8] that, like the the SM, also has an infinite state cardinality but, unlike the SM, is biased towards state reuse. State reuse results in an overall simpler latent mechanism. Unfortunately, hierarchical sharing of statistical strength is more difficult in the PDIA because the states in such a model lose their explicit correspondence with states in the world and it is unclear how such hierarchical sharing should be arranged.

Others have worked on "more complex" models, also reposed on infinite state spaces, that allow for non-deterministic transitions between states [3], non-deterministic transitions with side memory [7], and so on and so forth. The classical analogs of these models are given different names by different scientific communities. Computer scientists refer to these models as recognizers of formal languages in the Chomsky hierarchy [6]. Statisticians generally refer to them by the name of the corresponding generative model: Markov model, finite state automata, hidden Markov model, probabilistic context free grammar, etc. As a rule, in terms of classical complexity and expressivity, as one ascends this hierarchy the machines become more expensive to simulate, but more efficiently expressive on a per state basis. In our experiments with this spectrum of infinitely complex models we have noticed a profoundly perplexing pattern. For fixed-size data, the more complex (in the classical sense) models typically perform worse in general on real tasks. There are many reasons why this could be so (broken inference, dominance of a misspecified prior, relative importance of state reuse vs. hierarchical sharing, etc), but figuring out why this is so (when it should not be) is an important question we intend to address in this work.

The spectrum of models reposed on infinite state spaces are all hierarchical "Bayesian nonparametric" (BNP) models (the SM and PDIA form the bottom two rungs). BNP models can, in general, be made to conform to the AI modeling desiderata. They have uniformly exhibited excellent empirical performance on a wide variety of tasks, are endowed with the attractive properties of nonparametric estimators, and typically (via exchangeability properties) can be estimated efficiently and incrementally from streaming data. A variety of constant memory approximations to BNP models have been developed that also perform well. All of this suggests that BNP models should perform well as the models in which AI search and planning are performed. Whether this is true remains mostly unknown. This is the main hypothesis that we aim to test.

Summary

AI is a grand challenge. Solving AI requires developing models that can represent the state of the world in a task specific way that generalizes well using simple and easy to simulate latent mechanisms. Hierarchical BNP models exhibit many of the required characteristics, some of which have been highlighted here, more of which exist. While BNP models are certainly not the only such models in the mix (deep belief nets, etc.), their utility for AI search and planning remains significantly under-explored. Progress in BNP modeling has cross-cutting benefits; progress in BNP modeling for AI has the potential to produce very significant societal benefits.

Annual Budget

Travel	2 International conferences	$10,000
	2 Domestic conferences	$5,660
Salary	2 Post-doctoral fellows	$105,000
	PI summer (2 months)	$19,500
	Fringe 28.5%	$35,482
Total (salary)		$159,982
Total		$175,000

References

[1] 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.

[2] Bartlett, N. and Wood, F. (2011). Streaming deplump. In *Data Compression Conference 2011*.

[3] Beal, M., Ghahramani, Z., and Rasmussen, C. (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems*, 1:577–584.

[4] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer, New York, NY.

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

[6] Hopcroft, J., Motwani, R., and Ullman, J. (1979). *Introduction to automata theory, languages, and computation*. Addison-wesley Reading, MA.

[7] Johnson, M., Griffiths, T., and Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems*, volume 19, page 641.

[8] Pfau, D., Bartlett, N., and Wood, F. (2010). Probabilistic deterministic infinite automata. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1930–1938. MIT Press.

[9] Russell, S. J. and Norvig, P. (2002). *Artificial intelligence: A modern approach*. Prentice Hall, Englewood Cliffs, NJ, 2nd edition.

[10] 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.