Kangasrääsiö, Antti; Kaski, Samuel

Inference of Strategic Behavior based on Incomplete Observation Data

Published in:
NIPS17 Workshop: Learning in the Presence of Strategic Behavior

Published: 08/12/2017

Document Version
Peer reviewed version

Please cite the original version:
Kangasrääsiö, A., & Kaski, S. (2017). Inference of Strategic Behavior based on Incomplete Observation Data. In *NIPS17 Workshop: Learning in the Presence of Strategic Behavior* Carnegie Mellon University.

Inference of Strategic Behavior based on Incomplete Observation Data

Antti Kangasrääsiö
Department of Computer Science
Aalto University
antti.kangasraasio@aalto.fi

Samuel Kaski
Department of Computer Science
Aalto University
samuel.kaski@aalto.fi

Abstract

Inferring the goals, preferences and restrictions of strategically behaving agents is a common goal in many situations, and an important requirement for enabling computer systems to better model and understand human users. Inverse reinforcement learning (IRL) is one method for performing this kind of inference based on observations of the agent's behavior. However, traditional IRL methods are only applicable when the observations are in the form of state-action paths – an assumption which does not hold in many real-world modelling settings. This paper demonstrates that inference is possible even with an arbitrary observation noise model.

1 Introduction

Inverse reinforcement learning (IRL) [1, 2] is a problem where the reward function parameters of a Markov decision process (MDP) are inferred based on state-action observations of an optimally behaving agent.

A limitation with the traditional problem formulation is the assumption that full paths containing both actions and states have been observed. In many real-world situations such fine-grained observations may not be available for multiple reasons. For example, it may be costly to set up sensors that could gather the fine-grained observations, or it may be impossible to change the measurement devices if they are owned by a third party. Also, even if accurate sensors are used, various environmental factors may cause unavoidable occlusion, censoring or distortion to the measurements. Initial approaches have assumed that instead of the actual paths, we might only observe the feature expectations from the demonstrated paths [3], or alternatively that the state observations are probabilistic [4]. However, these two existing methods are not applicable when the observation noise model is more general.

This paper formulates the IRL from summary data (IRL-SD) problem, which extends the IRL problem to situations where the expert demonstrations are not available as complete state-action paths. We demonstrate that inference is possible with an arbitrary observation noise model, thus significantly extending the scope of problems where IRL can be performed. We derive the exact likelihood for this problem and two approximations. We demonstrate that all of these methods are able to solve a challenging toy problem, but the approximate methods scale significantly better. We demonstrate that a sensible approximate posterior can be inferred based for a recent RL-based model for the human oculomotor system based on incomplete observation data.

2 IRL from Summary Data

Let M be a MDP parameterized by θ and assume an agent whose behavior agrees with an optimal policy for M_{θ^*}. Assume that the agent has taken paths $(\xi_1, \ldots, \xi_N)$ but we only have observed

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

summaries of these paths: $\Xi_\sigma = (\xi_{1\sigma}, \ldots, \xi_{N\sigma})$, where $\xi_{i\sigma} \sim \sigma(\xi_i)$ and $\sigma(\xi_i) = P(\xi_{i\sigma}|\xi_i)$ is a stochastic summary function. The *inverse reinforcement learning problem from summary data (IRL-SD)* problem is:
Given (1) set of summaries Ξ_σ from optimal behavior; (2) summary function σ; (3) MDP M with θ unknown; and optionally (4) prior $P(\theta)$.
Determine $\hat{\theta}$ or the posterior $P(\theta|\Xi_\sigma)$.

3 Inference Methods for IRL-SD

To derive a computable likelihood, we assume both $|S|$ and $|A|$ to be finite and that the maximum number of actions within an observed episode is T_{max}. We denote the finite set of all plausible trajectories by $\Xi_{ap} \subseteq S^{T_{max}+1} \times A^{T_{max}}$.

The first alternative is to evaluate the true likelihood for θ given summary observations Ξ_σ, which is $L(\theta|\Xi_\sigma) = \prod_{i=1}^{N} \sum_{\xi_i \in \Xi_{ap}} P(\xi_{i\sigma}|\xi_i) P(\xi_i|\theta)$, where $P(\xi_{i\sigma}|\xi_i)$ is determined by the summary function σ, which is assumed to be known.

The second alternative is to use a Monte-Carlo estimate. In this approach, paths Ξ_{MC} (set of size $N_{MC} << |\Xi_{ap}|$) are simulated using policy π^*_θ, so that each path is drawn with probability $P(\xi|\theta)$. The likelihood of each individual observation can be estimated by a Monte-Carlo sum: $\hat{L}(\theta|\Xi_\sigma) = \prod_{i=1}^{N} \frac{1}{N_{MC}} \sum_{\xi_n \in \Xi_{MC}} P(\xi_{i\sigma}|\xi_n)$.

The third alternative is to avoid evaluating the likelihood function entirely, and use an approximate Bayesian computation (ABC) approach [5] instead. ABC also uses Monte-Carlo samples for estimating the likelihood, but does it by comparing the samples directly to the observation data using a *discrepancy function*, which is often chosen to be similar to the prediction error function. As this approach matches the overall behavior of the agent, it can also be seen as "IRL through imitation learning" [3]. The discrepancy function is denoted by $\delta(\Xi^A_\sigma, \Xi^B_\sigma) \to [0, \infty)$. Using δ we can define a stochastic variable $d_\theta \sim \delta(\Xi^{sim}_\sigma, \Xi_\sigma)$, where $\Xi^{sim}_\sigma = \{\sigma(\Xi_{MC,n})\}_{n=1\ldots|\Xi_\sigma|}$. The ability of θ to generate data similar to the observation data is quantified by the distribution of d_θ. An ABC approximation for the likelihood is $\tilde{L}_\varepsilon(\theta|\Xi_\sigma) = P(d_\theta \leq \varepsilon|\theta)$, with an approximation threshold $\varepsilon \in [0, \infty)$.

Recent work has shown the feasibility of Gaussian process (GP) [6] surrogates for expensive likelihoods [7], also in the ABC setting [8]. We use this approach as well for all of the three methods, as the likelihoods we work with are expensive to evaluate. The Bayesian optimization (BO) [9] sampling strategy is used for concentrating the samples so that high likelihood regions are well estimated.

4 Experiments

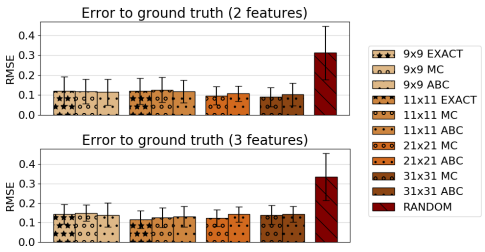
4.1 Experiment 1: Inference Quality

Our MDP here is a grid world environment. The agent is initially located on a random cell at the edge of the grid and wants to reach the center, but is blocked by different types of "soft walls". Our task is to infer how much each wall type hinders the agent. The summary function is defined as $\sigma(\xi) = (s_0, |\xi|)$, yielding the initial state at the edge, and the number of steps it took to reach the goal at center (i.e. we do not know what the intermediate states or actions were).

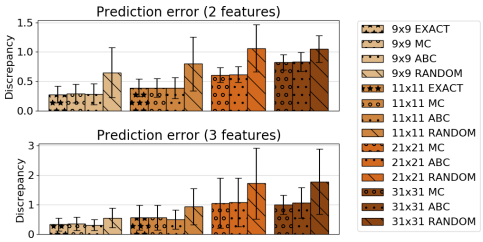
We measure inference quality both by accuracy of parameter recovery, which quantifies IRL performance, and prediction accuracy, which quantifies imitation learning performance. We observe that all of the methods are able to solve this problem equally well, and significantly better than a random baseline (Figures 1a and 1b). The approximate methods are able to perform well on larger grids where the exact method is computationally infeasible.

4.2 Experiment 2: Modelling Computer Users

We infer the full posterior of a recent RL-based cognitive model using realistic observation data. The task is to estimate the parameters of a MDP modelling the oculomotor system of a user who is searching for a specific menu-item from a computer drop-down menu [10, 11]. Although the state



(a) RMSE to ground truth (mean and standard deviation, N=30), smaller is better. NxN denotes grid size.



(b) Prediction error on test data (mean and standard deviation, N=30), smaller is better. NxN denotes grid size.

transition function is only defined as a computable algorithm, and the summary function σ is a delta distribution, the ABC method is still applicable.

We infer the posteriors of three parameters of the MDP: (1) duration of eye fixations f_{dur} (units of 100 ms); (2) duration of moving the mouse to select an item d_{sel} (units of 1 s); and (3) probability of recalling the full menu layout from memory p_{rec}. The posterior is visualized in Figure 2 using 2D slices at the MAP location ($d_{sel} = 0.05$, $p_{rec} = 0.80$, $f_{dur} = 2.6$). We observe that *a posteriori* there is correlation between f_{dur} and p_{rec}, and similarly for f_{dur} and d_{sel}. Both of these are understandable, as increasing f_{dur} would increase the predicted TCT, as would decreasing p_{rec} or increasing d_{sel}. The posterior of f_{dur} is centered around 260 ms, but there is still uncertainly left in d_{sel} and p_{rec}. The uncertainty in d_{sel} is explained by the difficulty of pointing precisely to the target item with the cursor, which causes variation in its duration. The uncertainty in p_{rec} is explained by the fact that the menus encountered early on in the experiments were completely new to the subjects, but as the experiment progressed the subjects were more and more likely to recall the menus. We also observe that there is no significant posterior correlation between p_{rec} and d_{sel}. This indicates that although they both affect the TCT, the effects they have are orthogonal; increasing the probability of recalling a menu can not be fully compensated just by increasing the selection duration.

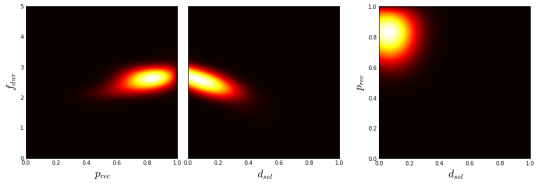


Figure 2: The approximate posterior inferred with ABC demonstrates that the parameters can be identified and that the remaining uncertainty is well characterized. Left: fixation duration f_{dur} and menu recall probability p_{rec}. Center: fixation duration f_{dur} and selection delay d_{sel}. Right: menu recall probability p_{rec} and selection delay d_{sel}. The color map is chosen so that the maximum of the posterior is white and minimum is black.

5 Discussion

Overall, regarding partial observability in IRL, there have been two cases for which methods exist: (1) If the agent has partial observability of the environment state, a POMDP model can be used [12]; (2) If the external observer has partial observability of the environment state, traditional IRL methods can be extended [4]. This work extends this list by a third item: (3) If the external observer has partial observability of the complete path, then the presented methods for IRL-SD can be applied.

Acknowledgements

This work has been supported by the Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, and grants 294238, 292334). Computational resources were provided by the Aalto Science IT project.

References

[1] Stuart Russell. Learning agents for uncertain environments. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, 1998.

[2] Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 663–670, 2000.

[3] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

[4] Kris M. Kitani, Brian D. Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Proceedings of the 12th European Conference on Computer Vision*, 2012.

[5] Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate Bayesian computation. *PLOS Computational Biology*, 9(1), 01 2013.

[6] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*. 2004.

[7] Carl Edward Rasmussen, JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals. In *Bayesian Statistics 7*, 2003.

[8] Michael U. Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(1), 2016.

[9] Eric Brochu, Vlad M. Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-023 and arXiv:1012.2599, University of British Columbia, 2010.

[10] Xiuli Chen, Gilles Bailly, Duncan P Brumby, Antti Oulasvirta, and Andrew Howes. The emergence of interactive behavior: A model of rational menu search. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[11] Antti Kangasrääsiö, Kumaripaba Athukorala, Andrew Howes, Jukka Corander, Samuel Kaski, and Antti Oulasvirta. Inferring cognitive models from data using approximate Bayesian computation. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems*, 2017.

[12] Jaedeug Choi and Kee-Eung Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12(Mar), 2011.