Kangasrääsiö, Antti; Kaski, Samuel

Modelling Human Decision-making based on Aggregate Observation Data

Published in:
Human In The Loop-ML Workshop at ICML

Published: 01/01/2017

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Kangasrääsiö, A., & Kaski, S. (2017). Modelling Human Decision-making based on Aggregate Observation Data. In *Human In The Loop-ML Workshop at ICML* Human in the Loop Machine Learning .

Modelling Human Decision-making based on Aggregate Observation Data

Antti Kangasrääsiö [1] **Samuel Kaski** [1]

Abstract

Being able to infer the goals, preferences and limitations of humans is of key importance in designing interactive systems. Reinforcement learning (RL) models are a promising direction of research, as they are able to model how the behavioural patterns of users emerge from the task and environment structure. One limitation with traditional inference methods for RL models is the strict requirements for observation data; both the states of the environment and the actions of the agent need to be observed at each step of the task. This has prevented RL models from being used in situations where such fine-grained observations are not available. In this extended abstract we present results from a recent study where we demonstrated how inference can be performed for RL models even when the observation data is significantly more coarse-grained. The idea is to solve the inverse reinforcement learning (IRL) problem using approximate Bayesian computation sped up with Bayesian optimization.

1. Introduction

Reinforcement learning (RL) based user models are based on the following assumptions: (1) there is an environment where the user is situated, (2) there is a task the user is trying to perform in this environment, (3) the user has performed similar tasks before and thus learned an optimal way to perform the task. A Markov decision process (MDP) is then constructed to match the situation, and a RL algorithm is used for solving the optimal behaviour policy. When the parameters of the MDP are unknown, inferring their values based on observations of the behavior of the user is known as inverse reinforcement learning (IRL).

Traditional methods for solving the IRL problem have been used in multiple real-world modelling situations, such as driver route modelling (Ziebart et al., 2008), helicopter acrobatics (Abbeel et al., 2010), learning to perform motor tasks (Boularias et al., 2011), dialogue systems (Chandramohan et al., 2011), pedestrian activity prediction (Ziebart et al., 2009; Kitani et al., 2012), and commuting routines (Banovic et al., 2016).

However, the traditional problem formulation assumes that we have observed the states of the environment and the actions of the agent at each step of the decision process. In many real-world situations it may not be feasible to collect such fine-grained observations, while other types of observations may be easily available. For example, (1) we may be limited by budget from collecting fine-grained observations, (2) we may be in an adversarial situation, where the opponent prevents us from making accurate observations, (3) we may have privacy related reasons that restrict our access to accurate observations.

To extend the applicability of RL models, we propose a variant of the IRL problem, called the inverse reinforcement learning from summary data (IRL-SD) problem. We then propose two methods for solving the problem: one based on the exact observation likelihood and another based on an approximate likelihood, inspired by approximate Bayesian computation (ABC). For inference we propose a Bayesian optimization (BO) based method.

We demonstrate that use of the surrogate allows us to infer both maximum likelihood estimates and full posteriors of parameters on moderate-sized models, including a cognitive model for human visual search, based on only aggregate observations. Further details are presented in the original publication (Kangasrääsiö & Kaski, 2017).

2. IRL from Summary Data

2.1. Problem Definition

Let M be a MDP (S, A, T, R, γ) with parameters θ. Let the true parameters be $\theta^* \in \Theta$ and assume agent behaving according to an optimal policy for M_{θ^*}. Assume the agent has taken paths $(\xi_1, \ldots, \xi_N)$ and we observe *summaries* $\Xi_\sigma = (\xi_{1\sigma}, \ldots, \xi_{N\sigma})$, where $\xi_{i\sigma} \sim \sigma(\xi_i)$ and σ is a known summary function. The *inverse reinforcement learning problem from summary data (IRL-SD)* is then:

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland. Correspondence to: Antti Kangasrääsiö <antti.kangasraasio@aalto.fi>, Samuel Kaski <samuel.kaski@aalto.fi>.

Given (1) set of summaries Ξ_σ of an agent demonstrating optimal behavior; (2) summary function σ; (3) MDP M with θ unknown; (4) bounded space Θ; and optionally (5) prior $P(\theta)$.
Estimate $\hat{\theta} \in \Theta$ such that simulated behavior from $M_{\hat{\theta}}$ agrees with Ξ_σ, or the posterior $P(\theta|\Xi_\sigma)$.

2.2. Exact Likelihood

Assume both $|S|$ and $|A|$ are finite and that the maximum number of actions that can be performed within an observed episode is T_{max}. Denote the finite set of all plausible trajectories by $\Xi_{ap} \subseteq S^{T_{max}+1} \times A^{T_{max}}$.

The likelihood for θ given $\Xi_\sigma = (\xi_{1\sigma}, \ldots, \xi_{N\sigma})$ is now

$$L(\theta|\Xi_\sigma) = \prod_{i=1}^{N} \big[P(\xi_{i\sigma}|\theta)\big] = \prod_{i=1}^{N} \Big[\sum_{\xi_i \in \Xi_{ap}} \big[P(\xi_{i\sigma}|\xi_i)P(\xi_i|\theta)\big]\Big],$$

where

$$P(\xi_{i\sigma}|\xi_i) = P\big(\sigma(\xi_i) = \xi_{i\sigma}\big),$$

and

$$P(\xi_i|\theta) = P(s_0^i) \prod_{t=0}^{T_i-1} \big[\pi_\theta^*(s_t^i, a_t^i) P(s_{t+1}^i|s_t^i, a_t^i)\big].$$

2.3. Approximate Likelihood

Assume a function for generating summary datasets Ξ_σ^{sim} given MDP M, parameters θ, number of episodes N, and summary function σ: $\texttt{RLSUM}(M_\theta, N, \sigma)$. Also assume a *discrepancy function* δ,

$$\delta(\Xi_\sigma^A, \Xi_\sigma^B) \to [0, \infty),$$

which quantifies the dissimilarity between two observation datasets. Note that δ needs to be selected based on domain knowledge, or learned from data (Gutmann et al., 2017). By combining $\texttt{RLSUM}(M_\theta, |\Xi_\sigma|, \sigma)$ with δ, we define

$$d_\theta \sim \delta(\texttt{RLSUM}(M_\theta, |\Xi_\sigma|, \sigma), \Xi_\sigma).$$

The distribution of d_θ corresponds with the ability of θ to satisfy our requirements for solving the IRL-SD problem. Finally we define an approximate likelihood function,

$$\tilde{L}_\varepsilon(\theta|\Xi_\sigma) = P(d_\theta \leq \varepsilon|\theta),$$

where the approximation threshold $\varepsilon \in [0, \infty)$. The threshold ε can be chosen adaptively (Lintusaari et al., 2017).

2.4. Inference

As computing the gradients is not feasible, and evaluating even the surrogate is expensive, we use Bayesian optimization (BO) (see Brochu et al., 2010) combined with a Gaussian process (GP) regression model (Rasmussen, 2004) for finding the maximum likelihood (ML) point estimate. For a GP fit with data D and hyperparameters H, the mean at θ is $G_\mu(\theta|D, H)$ and the standard deviation is $G_\sigma(\theta|D, H)$. The optimization horizon of BO is N_{opt} and the acquisition function value at θ is $Acq(\theta|D, H)$ (the maximum of the acquisition function defines the next sample location in BO).

Algorithm 1 summarizes exact ML inference and Algorithm 2 the approximate (we use the discrepancy values instead of approximate likelihood values for convenience).

Algorithm 1 Exact Maximum Likelihood Inference Algorithm for IRL-SD

Input: M, Ξ_σ, Θ, H, N_{opt}
Output: $\hat{\theta}_{ML}$
$D \leftarrow \varnothing$
for $i = 1$ **to** N_{opt} **do**
  $\theta_i \leftarrow \arg\max_\theta Acq(\theta|D, H)$
  $\pi_{\theta_i}^* \leftarrow \texttt{RL}(M_{\theta_i})$
  $l_\theta \leftarrow -\log L(\theta_i|\Xi_\sigma)$
  $D \leftarrow \{D, (\theta_i, l_\theta)\}$
end for
$\hat{\theta}_{ML} \leftarrow \arg\min_\theta G_\mu(\theta|D, H)$

Algorithm 2 Approximate Maximum Likelihood Inference Algorithm for IRL-SD

Input: M, Ξ_σ, Θ, H, N_{opt}
Output: $\hat{\theta}_{ML}$
$D \leftarrow \varnothing$
for $i = 1$ **to** N_{opt} **do**
  $\theta_i \leftarrow \arg\max_\theta Acq(\theta|D, H)$
  $\Xi_\sigma^{sim} \leftarrow \texttt{RLSUM}(M_{\theta_i})$
  $d_\theta \leftarrow \delta(\Xi_\sigma^{sim}, \Xi_\sigma)$
  $D \leftarrow \{D, (\theta_i, d_\theta)\}$
end for
$\hat{\theta}_{ML} \leftarrow \arg\min_\theta G_\mu(\theta|D, H)$

For exact posterior inference, the log-likelihood in Algorithm 1 can be replaced with log-posterior. For approximate inference we take a similar approach as Gutmann and Corander (Gutmann & Corander, 2016) by returning $\tilde{P}(\theta|\Xi_\sigma) = P(\theta)\tilde{L}_\varepsilon(\theta|\Xi_\sigma)$ with the distribution of d_θ estimated from the GP.

3. Experiments

3.1. Grid World

As a toy example we used a variation of the grid world problem. In our version the agent was initially placed at the edge of a square grid and the task of the agent was to get to

the center of the grid. Our task was to infer the parameters of a linear reward function for the features of the states, $R(s) = \phi(s)^T\theta$, where $\phi(s)$ is the binary feature vector of state s. Our summary observation function σ extracted the initial location of the agent and the length of the episode.

We first estimated the runtime of the algorithms as a function of the grid size (Fig. 1), and noticed that the exact method is not computationally feasible for large grids. We estimated the quality of the inference on various sizes of grids (Fig. 2) using both the prediction error (measured with the discrepancy function δ) and inference quality (L_2 error for ML estimates). We observed that both methods get results better than random, and that the approximate method actually performs better than the exact one on this problem. We suspect that this is because matching the global features of the behavior is likely more robust than matching the local state-transition probabilities in the likelihood function.

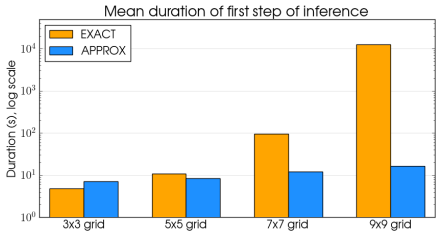

Figure 1. Duration of the first step of the exact and approximate ML estimation algorithms (mean of 20 experiments).

3.2. Experiment 3: Modelling Computer Users

We also inferred the full posterior of a cognitive science model with the approximate method based on real observation data. The MDP models a user performing visual search from a computer drop-down menu (Chen et al., 2015; Kangasrääsiö et al., 2017). With small computer menus the accuracy of eye-tracking is often poor, so it is difficult to get reliable measurements at the state-action level. However, simple summary statistics, such as the time between opening a menu and clicking the target item, are simple to measure accurately.

Our summary observation included the task completion time (TCT) and whether the target was present or absent in the menu, and the discrepancy is based on the differences in TCT distributions. We infer the posteriors of three parameters: (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}. To make visualization easier, the posterior is inferred in two groups: first for f_{dur} and p_{rec} (Figure 3 left), then for f_{dur} and d_{sel} (Figure 3 right).

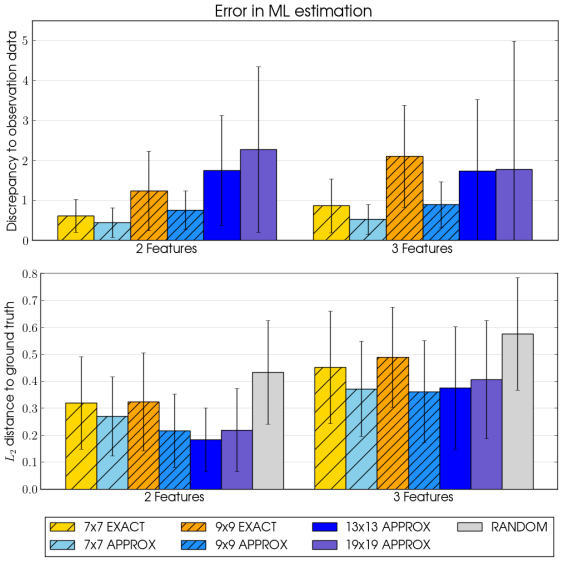

Figure 2. Quality of ML estimates with the exact and approximate inference methods. Top: discrepancy to the observation data, smaller is better. Bottom: L_2 distance to ground truth, smaller is better. Random baseline represents a uniform guess from the parameter space. The bars show the mean and standard deviation of 30 independent experiments.

We observe a correlation between f_{dur} and p_{rec}, and similarly for f_{dur} and d_{sel}. These are understandable, as increasing f_{dur} increases the TCT, as would decreasing p_{rec} or increasing d_{sel}. There is still considerable uncertainly left in d_{sel} and in p_{rec}. The uncertainty in d_{sel} could be explained by individual variation in selection times, and p_{rec} by the fact that the menus encountered early on in the experiments were completely new to the subjects, but at the end of the experiment the subjects were more likely to recall a previously encountered menu. We note that these insights would not have been possible to infer from just MAP estimates.

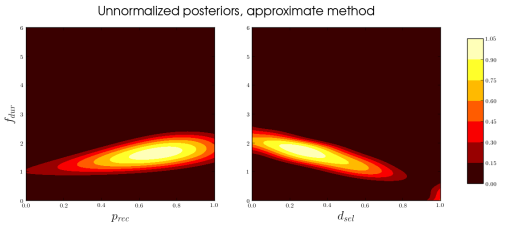

Figure 3. Left: posterior of f_{dur} and d_{sel}. Right: posterior of f_{dur} and p_{rec}.

4. Discussion

We defined the IRL-SD problem, proposed exact and approximate methods for inference and demonstrated that

they both are able to solve the inference problem. We demonstrated that the approximate method is scalable enough to be used for full posterior estimation for a realistic cognitive science model.

Regarding partial observability in IRL, there now exists formulations for three different situations. (1) If the agent has partial observability of the environment state, a POMDP model can be used (Choi & Kim, 2011). (2) If the external observer has partial observability on environment state level, traditional IRL methods can be extended (Kitani et al., 2012). (3) If the external observer has partial observability on complete path level, 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

Abbeel, Pieter, Coates, Adam, and Ng, Andrew Y. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research*, 29(13):1608–1639, 2010. doi: 10.1177/0278364910371999.

Banovic, Nikola, Buzali, Tofi, Chevalier, Fanny, Mankoff, Jennifer, and Dey, Anind K. Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 248–260, 2016. doi: 10.1145/2858036.2858557.

Boularias, Abdeslam, Kober, Jens, and Peters, Jan. Relative entropy inverse reinforcement learning. In *JMLR Workshop and Conference Proceedings Volume 15: AISTATS 2011*, pp. 182–189, 2011.

Brochu, Eric, Cora, Vlad M., and De Freitas, Nando. 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.

Chandramohan, Senthilkumar, Geist, Matthieu, Lefevre, Fabrice, and Pietquin, Olivier. User simulation in dialogue systems using inverse reinforcement learning. In *Interspeech*, pp. 1025–1028, 2011.

Chen, Xiuli, Bailly, Gilles, Brumby, Duncan P, Oulasvirta, Antti, and Howes, Andrew. 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*, pp. 4217–4226, 2015. doi: 10.1145/2702123.2702483.

Choi, Jaedeug and Kim, Kee-Eung. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12(Mar):691–730, 2011.

Gutmann, Michael U. and Corander, Jukka. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(1):4256–4302, 2016.

Gutmann, Michael U., Dutta, Ritabrata, Kaski, Samuel, and Corander, Jukka. Statistical inference of intractable generative models via classification. *Statistics and Computing*, 2017. In press, preprint arXiv:1407.4981.

Kangasrääsiö, Antti and Kaski, Samuel. Inverse reinforcement learning from summary data. *arXiv preprint arXiv:1703.09700*, 2017.

Kangasrääsiö, Antti, Athukorala, Kumaripaba, Howes, Andrew, Corander, Jukka, Kaski, Samuel, and Oulasvirta, Antti. Inferring cognitive models from data using approximate Bayesian computation. In *Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems*, pp. 1295–1306, 2017. doi: 10.1145/3025453.3025576.

Kitani, Kris M., Ziebart, Brian D., Bagnell, James Andrew, and Hebert, Martial. Activity forecasting. In *Proceedings of the 12th European Conference on Computer Vision*, pp. 201–214, 2012. doi: 10.1007/978-3-642-33765-9_15.

Lintusaari, Jarno, Gutmann, Michael U., Dutta, Ritabrata, Kaski, Samuel, and Corander, Jukka. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):e66, 2017. doi: 10.1093/sysbio/syw077.

Rasmussen, Carl Edward. Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning*, pp. 63–71. 2004. doi: 10.1007/978-3-540-28650-9_4.

Ziebart, Brian D., Maas, Andrew L., Bagnell, J. Andrew, and Dey, Anind K. Maximum entropy inverse reinforcement learning. In *Proceedings of the National Conference on Artificial Intelligence*, pp. 1433–1438, 2008.

Ziebart, Brian D., Ratliff, Nathan, Gallagher, Garratt, Mertz, Christoph, Peterson, Kevin, Bagnell, J. Andrew, Hebert, Martial, Dey, Anind K., and Srinivasa, Siddhartha. Planning-based prediction for pedestrians. In *International Conference on Intelligent Robots and Systems*, pp. 3931–3936, 2009. doi: 10.1109/IROS.2009.5354147.