Learn and Represent Human Utility via Preference

Feiyang Xie
Yuanpei College
Peking University
2100017837@stu.pku.edu.cn

Abstract

Rooted in the intersection of philosophy, economics, and game theory, the concept of utility stands as a foundational principle within modern decision theory. In this framework, an agent's rational decision-making is guided by the pursuit of maximizing its expected utility, encapsulated by the principle of maximum expected utility. In the realm of daily life, individuals not only shape their current behaviors based on utility functions but also employ them to decipher the actions and intentions of others, as well as discern cause-and-effect relationships within their environment. The acquisition and representation of human utility have emerged as a pivotal research area, particularly within fields such as computer vision (CV) and reinforcement learning (RL). Given the intricate nature of human utility functions, direct modeling proves challenging. Consequently, the predominant approach involves learning from human preferences. In the context of RL, the assimilation of human preferences is often conceptualized as the acquisition of a reward function, aligning with the agent's overarching objective to maximize cumulative rewards. This essay endeavors to explore the representation of human preference, delving into various computational frameworks for utility learning while scrutinizing their respective advantages and disadvantages. Subsequently, it will delve into the intricate landscape of human utility within large language models (LLMs) and elucidate the limitations inherent in preference-based methodologies.

1 Introduction

Utility, often indicative of preference in decision-making, stands as a fundamental tenet in modern decision theory, encapsulated by the principle of maximum expected utility [1]. Tasking agents with learning the utility function of humans holds significant import, facilitating the alignment of agent behavior with human preferences and thereby enhancing the agent's capacity to comprehend and aid humans in task completion. The inherent diversity and complexity of individual human preferences present a formidable challenge in directly modeling human utility functions. A one-size-fits-all framework risks introducing bias, potentially leading to misaligned behaviors. To mitigate this bias, contemporary approaches frequently leverage human preferences as a basis for modeling utility functions. This involves utilizing preferences and feedback to guide the training of models. In reinforcement learning (RL), for instance, human preferences often influence the reward function, serving as a surrogate for the utility function. Given the agent's overarching objective to maximize cumulative rewards, this approach incrementally refines the agent's behavior to align with human expectations.

This paper commences by exploring extant representations of human preferences and subsequently introduces computational frameworks for utility learning. Simultaneously, it scrutinizes the strengths and weaknesses of these frameworks and representation methods, considering factors such as data collection, generalization, and efficiency. Finally, the discourse extends to the domain of human utility learning within large language models (LLMs), elucidating strategies to enhance the alignment between LLMs and the intricate contours of human utility functions.

2 The representation of preference

Human preferences are the external manifestation of utility functions. Preferences are more subjective and individual than objective values. Given a specific scenario, a decision based on preference may not have the highest objective value, but it must be the best subjectively for the individual. This section mainly discusses the representation of preferences in RL.

In RL, the evaluation objects of preference are usually the following:

1. **Action-only**: Given an initial state s_0 and series of actions $(a_0, a_1, ...a_t)$, humans select better outputs based on predictions of environmental changes.
2. **State-action pair**: Given some state-action pairs $(s_t, a_t^0), (s_t, a_t^1), ..., (s_t, a_t^k)$, a preference $(s_t, a_t^i) \prec (s_t, a_t^j)$ denotes $\hat{r}(s_t, a_t^i) \leq \hat{r}(s_t, a_t^j)$, where $\hat{r}$ indicates human utility function on this environment.
3. **Segment of trajectory**: A trajectory segment is a sequence of observations and actions, $\sigma = ((s_1, a_1), (s_1, a_1), ..., (s_k, a_k))$. Write $\sigma^1 \prec \sigma^2$ to indicate that the human preferred trajectory segment σ^2 to σ^1.

Due to the fuzziness of preference itself, related research has also diversified the way it affects the training process. Here we briefly introduce some simpler and widely used modeling methods. The specific methods are presented in the table 1 [2].

Preference	Constraint	Probablistic
Comparisons	$r(\xi_1) \geq r(\xi_2)$	$\mathbb{P}(\xi_1 \vert r) = \frac{exp(\beta \cdot r(\xi_1))}{exp(\beta \cdot r(\xi_1)) + exp(\beta \cdot r(\xi_2))}$
Demonstrations	$r(\xi_D) \geq r(\xi), \forall \xi$	$\mathbb{P}(\xi_1 \vert r) = \frac{exp(\beta \cdot r(\xi_1))}{exp(\beta \cdot \xi_D) + exp(\beta \cdot \xi_2)}$
Reward and Punishment	$r(\xi_R) \geq r(\xi_{expected})$	$\mathbb{P}(+1 \vert r) = \frac{exp(\beta \cdot r(\xi_R))}{exp(\beta \cdot r(\xi_R)) + exp(\beta \cdot r(\xi_e xpected))}$

Table 1: The probabilistic modeling of preferences

Comparisons Comparison is the most common and intuitive way. In comparisons, the human is typically shown two trajectories and then asked to select the one that they prefer. Human choice space is two trajectories ξ_1, ξ_2. Most work on comparisons is done in the preference-based RL domain in which the robot might compute a policy directly to agree with the comparisons, rather than explicitly recover the reward function [3].

Demonstrations In demonstrations, the human is asked to demonstrate the optimal behavior. Learning reward function from demonstrations is often called inverse reinforcement learning (IRL) [4]. A commonly used assumption is that human decisions are optimal in any situation. However, as the complexity of the environment increases, the difficulty of the task increases, and the action spaces faced are different, the strategies adopted by humans may fall into sub-optimal.

Reward and Punishment In this type of feedback, the human can either reward $+1$ or punish -1 the robot for its trajectory. It's obvious that humans reward and punish based on how well the robot performs relative to their expectations. So, we have an human expected trajectory $\xi_{expected}$ and if ξ_R is better than $\xi_{expected}$, we will give it a reward $+1$. This is also the biggest difference between Reward and Punishment and Comparisons.

3 Computation frameworks for utility learning

3.1 Traditional RL framework

Based on the form of modeling preferences discussed above, a lot of related work try to fit and learn reward function based on human preferences. After obtaining a value function that is aligned with humans, the strategy can be optimized well using the traditional RL framework. For example, in [5], the preferences is modeled as the following:

$$\hat{P}(\sigma^1 \prec \sigma^2) = \frac{exp \sum \hat{r}(o_t^2, a_t^2)}{exp \sum \hat{r}(o_t^1, a_t^1) + exp \sum \hat{r}(o_t^2, a_t^2)}$$

$$loss(\hat{r}) = -\sum \mu(1) log\hat{P}(\sigma^2 \prec \sigma^1) + \mu(2) log\hat{P}(\sigma^1 \prec \sigma^2)$$

where $\mu()$ refers to the choice of human.

This paradigm is widely adopted in preference-based RL, but it usually requires a large amount of human-labeled data. And due to the singleness of the training target, learning the reward function can easily lead to reward hacking, causing deviations.

3.2 ZO-RankSGD

Black-box optimization, which utilizes a derivative-free framework to optimize the target function,has been extensively studied in the optimization literature for several decades. In some cases, it is difficult for humans to quantify preferences and we try to let the model learn the human utility function when the model parameters are unknown. ZO-RankSGD is an effective approach for model optimization that finds the descent direction directly from the ranking information [6]. Given a function $f : \mathbb{R}^d \to \mathbb{R}$ and m points $x_1, ..., x_m$ to query, an (m, k) ranking oracle $O_f^{(m,k)}$ will return k smallest points sorted in their order. Given a start point x, we can query $O_f^{(m,k)}$ with the inputs $x_i = x + \mu\xi_i, \xi_i \sim N(0, I_d)$, for $i = 1, ..., m$. We can construct a directed acyclic graph $G = (N, \epsilon)$, where the directed edge set $\epsilon = (i, j)|f(x_i) < f(x_j)$, the rank-based gradient estimator can be formulated as:

$$\tilde{g}(x) = \frac{1}{|\epsilon|} \sum_{(i,j)\in\epsilon} \frac{x_j - x_i}{\mu} = \frac{1}{|\epsilon|} \sum_{(i,j)\in\epsilon} (\xi_j - \xi_i)$$

With this method we can try to introduce human preferences to some pre-trained large models, such as prompt-tuning for in-context learning. This method is suitable for most downstream tasks, and it is usually more efficient to introduce preferences because it only modifies a small number of parameters. In terms of data collection, this approach requires humans to rank various choices without giving precise values.

4 Human utility learning in LLM

In recent times, large language models like GPT-4 have demonstrated remarkable human-like capabilities across various domains. However, if left unguided during the training phase, these models can exhibit misaligned behaviors, including tendencies towards Power-Seeking Behaviors and Hallucination, as noted by Ji et al. [7]. Addressing this challenge requires the incorporation of human preferences and utility functions into the model. One promising approach to mitigate misalignment is Reinforcement Learning from Human Feedback (RLHF). In this paradigm, human evaluators provide feedback by comparing alternative responses generated by the trained conversation model. The collected feedback is then utilized through reinforcement learning against a pre-trained reward model. Nevertheless, a significant hurdle in this context is achieving scalable oversight, particularly when dealing with super-human AI systems operating in complex scenarios beyond the comprehension of human evaluators. In such situations, the behaviors of these systems may prove challenging for humans to comprehend and evaluate, raising concerns about the quality of feedback [8].

Our overarching objective is to enable models to learn a universal utility function. In essence, the choices made by the model, based on this function, should consistently align with human utility across diverse situations. However, the current landscape poses difficulties in defining utility functions for problems beyond human-solving capabilities. Additionally, models struggle to learn and represent utility functions through causal reasoning in such contexts. This underscores the need for innovative solutions to bridge the gap between model decision-making and complex, non-human-solvable problems.

5 Conclusion

As the foundational underpinning of human decision-making, the acquisition and representation of utility functions stand as indispensable stages in the development of more sophisticated intelligent agents. Presently, utility functions are predominantly modeled and learned through preferences, with preference serving as a mapping outcome derived from the human utility function. However, it falls short of fully encapsulating the intricacies of the utility function.

While evaluating outcomes and expressing preferences is ostensibly simpler than the demonstration process, the evolution of AI introduces complexities, demanding a substantial depth of professional knowledge for result assessment. This complication, in turn, poses challenges in data collection. Consequently, a promising avenue for future research lies in the pursuit of more efficient and effective methods for modeling utility functions. Additionally, there is a pressing need to devise streamlined approaches for conveying utility function information that aligns more seamlessly with human understanding. Addressing these challenges will pave the way for the creation of intelligent agents capable of comprehending and aligning with human utility in a manner that is both nuanced and accessible.

References

[1] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 1

[2] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020. 2

[3] Aaron Wilson, Alan Fern, and Prasad Tadepalli. A bayesian approach for policy learning from trajectory preference queries. *Advances in neural information processing systems*, 25, 2012. 2

[4] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000. 2

[5] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 2

[6] El Houcine Bergou, Eduard Gorbunov, and Peter Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM Journal on Optimization*, 30(4):2726–2749, 2020. 3

[7] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 3

[8] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022. 3