Possible ways to learn and represent human utility

Xingping Yu
Peking University
2100017812@stu.pku.edu.cn

Abstract

Utility function plays an important role in human decision making. But these utility functions are internal to humans, which are hard to observe and represent in a specific way. Moreover, utility functions can vary from different individuals. In another word they are quite subjective and hard to measure with a consistent standard. When it comes to different tasks, there are more kinds of utility functions behind the decision-making process. For computational models, we are trying to build some general frameworks to learn and represent the utility functions that are close to human utility. Some possible ways are discussed in this essay. This essay mainly proposes some possible utility functions like cost, human preference etc and learning policies like deep learning and reinforcement learning. For further analysis, this essay discusses some advantages and disadvantages of the ways in data collection, generalization and efficiency.

1 Introduction

Human utility is a concept that describes an individual's preferences and satisfaction with different choices or actions. It is used to explain and predict people's decision-making behavior. An agent makes rational decisions or choices based on their beliefs and desires to maximize its expected utility, which is known as the principle of maximum expected utility[7] . However, these utility functions are different across individuals. Also, because the utility is related to the downstream task, utility functions are various and hard to measure with a computational model. To build computational framework, we should find ways to imitate human utility as possible as we can. An intuitive idea is that we can train the model to acquire values similar to humans. Some possible ways may include deep learning and reinforcement learning with proper reward functions.

2 Possible ways to represent human utility

To find possible ways to learn and represent human utility, we can first study the decision-making process of humans. Classic models in social choice theory assume that the preferences of a set of agents over a set of alternatives are represented as linear orders; a social choice function outputs a single socially desirable alternative given these preferences as input [1] . So an agent will maximize the function when it makes choices/ decisions. Inspired by existing works, there are some possible ways to represent these utility functions in my opinion.

2.1 Multi-factor average function

One naive way is to represent the utility function as a weighted average of multiple factors. This inspiration is actually from the work mentioned just now (Boutilier et al., 2012 [1]). Although that work focused on social choice function and multi-agent choice-making, I think we can replace the agents with different factors we consider when making decisions. In another words we can list a set of factors that influence our preferences when making decision. These factors will be weighted as parameters. And the final utility is the weighted average of these factors.

It seems that there is a linear mapping from these factors to the utility function. And we calculate utility of each candidate and choose one that maximize the function. This decision-making process is quite simple. But is that all we need?

2.2 Cost and risk function

Sometimes we don't measure a decision in terms of returns or the returns are not computable in a short term. For example, during the process we raise our kids, the returns and rewards are quite invisible, but we still spend a lot because we believe this process needs much cost. In another case, when we are investing in a project, the rewards are unpredictable sometimes. So we should consider the cost and risk rationally.

Study shows that ten-month-old infants infer the value of goals from the costs of actions [3] . This phenomenon is actually not out of expect. The infants can't completely understand the utility of goals. But they can measure the costs of actions through observation. And computational models can also adopt this kind of metric. A naive thought is that the more an action costs, the more utility it has. However, the cost and reward are not always positively correlated. So we need to add risk functions to modify the utility function. The final utility should be calculated by the cost function and risk function.

2.3 Human preference

The two ways to represent utility functions are both fixed. They try to simulate human preference from different aspects. But can we directly learn from human preference? This is an interesting topic recently. Actually, there are many existing works that focused on learning from human preference.

In a survey of preference-based reinforcement learning, preference-based reinforcement learning algorithms (PbRL) have been proposed that can directly learn from an expert's preferences instead of a hand-designed numeric reward [6] . It provided a unified framework for PbRL that describes the task formally and pointed out the different design principles that affect the evaluation task for humans. This work is quite inspiring because it adds human preference and evaluation directly into computational models. It makes it possible for models to learn human preferences in a direct way.

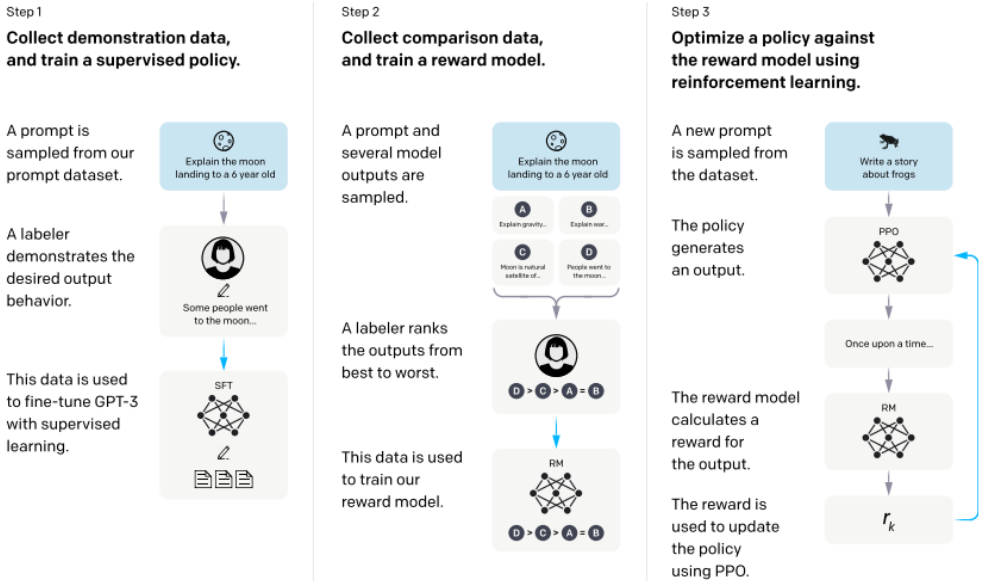

Figure 1: A diagram illustrating the three steps of InstructGPT training: (1) supervised fine-tuning (SFT), (2) reward model(RM) training, and (3) reinforcement learning via proximal policy optimization(PPO) on this reward model.

Based on this idea, reinforcement learning from human feedback (RLHF) is proposed for large language models like ChatGPT and InstructGPT [2, 5] . The core idea is to learn a reward model

(RM) from human feedback. Then use this reward model to do reinforcement learning. The current language models based on this technology perform very well in many downstream tasks like question answering and dialogue generation.

Take InstructGPT that was published in 2022 for example [4] . It scaled up RLHF to tens of thousands of tasks. The training process is represented in Figure 1. In the first step, it collects demonstration data and trains a supervised policy. This includes pretraining and supervised fine-tuning. In the second step, it collects comparison data and trains a reward model. This is the core step we care. There are certain metrics for human to evaluate and give feedback to model. The model will then learn the human preferences based on the data. They use comparison data instead of absolute data to decrease the differences between individuals. Finally the model optimizes a policy against the reward model using reinforcement learning. Given that models based on RLHF have made good progress in learning human utility, this method is quite available and promising.

3 Learning methods and their advantages/disadvantages

So far we have talked about several ways to represent human utility with computational framework, it's easy to think of some learning methods based on these representations. I will mainly discuss deep learning and reinforcement learning as examples.

Given the parameters of utility function, we can use deep learning to fit the model. For example, we adopt the multi-factor average function mentioned above as utility function. Then we can design a neural network and train it on the human-designed dataset. An obvious advantage of this model is that it is quite simple.The cost of data collection is small and the efficiency can be high. However, this method depends much on the data it receives, which means low generalization and transfer ability. Because each person has his/her own sets and weights of utility factors, it is hard to fit the parameters with a unified model.

Another way to learn human preference is reinforcement learning. The core step is to build a proper reward model or reward function. One possible way is to use cost and risk functions. We can add these functions to loss and reward functions. This method is easy to collect data. The dataset is similar to that of deep learning with multi-factor average function. It just need to pre-process the data before training. But the disadvantage of this method is also similar to multi-factor method. These cost and risk functions are static and hard to change. And the model will only fit the designed reward functions. However, the cost and risk functions can not represent human utility completely. Many times we consider costs and returns together to get utility. And even if the returns are sometimes invisible, we prefer to predict in our own way, which can be very internal. So the efficiency and accuracy of this method may be not satisfying.

So what about reinforcement learning with human feedback? The model will first learn human preferences from human feedback, and train with the reward function. It seems that this method is very ideal and can solve the problem. It indeed has made quite inspiring progress currently. The performance of the models is great in some downstream tasks. There are some evidences that prove the successful learning of human preferences to some degree. However, it is not the end of the problem. There are still many difficulties in front. For example, in terms of data collection, getting human evaluation and feedback is expensive and slow. And that's even not the main problem. Because utility functions are different across individuals, we still haven't solved the problem that the data can be too subjective. Although scaling up the dataset may decrease this effect, the data collection is always a problem. Moreover, the reinforcement learning model will find the easiest way to maximize reward function. So it will sometimes come to unexpected outputs even if the outputs maximize the reward function. In practice, we should mitigate these "shortcuts" or we hope that's aligned with the behavior we want.

In fact, the subjective problem is universal in the methods above. Because human utility itself can be quite complex and various. Learning and representing human utility based on simple dataset is never a easy task. Sometimes it can even learn some harmful utility functions that may raise moral and ethical questions. For example, Chatbot released by Microsoft in 2016 started making toxic racist and sexist comments within 24 hours. We need to find more proper ways to collect and process the data.

4 Conclusion

In conclusion, there are some possible ways to learn and represent human utility. We can use some factors and parameters to calculate utility. Also we can let the model directly learn from human preferences and adopt reinforcement learning with reward model. The models based on RLHF perform well in some downstream tasks currently. No matter what method we adopt, the data collection is a problem. The difference between individuals will influence the reward model. And the model should be modified properly based on different human preferences. In the future, we may scale up the dataset to learn more general utility function models. At the same time, we may improve the learning and representing methods to better fit human utility.

References

[1] Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 197–214, 2012. 1

[2] 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

[3] Shari Liu, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017. 2

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[5] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020. 2

[6] Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017. 2

[7] 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