Reward Elicitation Report

Soheil

July 3, 2018

Contents

Reward Elicitation

Reward quantification is a hard job, cognitively complex in practice, and time consuming. It is problem dependent, and when it comes to deciding a value for quantities like “good” or “mediocre”, user tend to not have a clear answer. In here, we try to obtain the optimal policy while having partial reward information.

What is preference elicitation problem?

The key idea is called minimax regret decision criterion, which is intuitively based on minimum regret or loss.

Agent-Based Modeling and Multiagent Systems

The **minimax regret** approach is to minimize the worst-case regret One benefit of minimax (as opposed to expected regret) is that it is independent of the probabilities of the various outcomes: thus if regret can be accurately computed, one can reliably use minimax regret. However, probabilities of outcomes are hard to estimate.

Minimax Regret Example from wikipedia

Suppose an investor has to choose between investing in stocks, bonds or the money market, and the total return depends on what happens to interest rates. The following table shows some possible returns:

Return	Interest rates rise	Static rates	Interest rates fall	Worst return
Stocks	-4	4	12	-4
Bonds	-2	3	8	-2
Money market	3	2	1	1
Best return	3	4	12	

The crude maximin choice based on returns would be to invest in the money market, ensuring a return of at least 1. However, if interest rates fell then the regret associated with this choice would be large. This would be 11, which is the difference between the 12 which could have been received if the outcome had been known in advance and the 1 received. A mixed portfolio of about 11.1% in stocks and 88.9% in the money market would have ensured a return of at least 2.22; but, if interest rates fell, there would be a regret of about 9.78.

The regret table for this example, constructed by subtracting actual returns from best returns, is as follows:

Regret	Interest rates rise	Static rates	Interest rates fall	Worst regret
Stocks	7	0	0	7
Bonds	5	1	4	5

Regret	Interest rates rise	Static rates	Interest rates fall	Worst regret
Money market	0	2	11	11

Therefore, using a minimax choice based on regret, the best course would be to invest in bonds, ensuring a regret of no worse than 5. A mixed investment portfolio would do even better: 61.1% invested in stocks, and 38.9% in the money market would produce a regret no worse than about 4.28.

Minimax

When dealing with multiple states and actions, sometimes, we want to *maximize* our *worst case*(min) in terms of rewards. We don't care what we are going to get when things were fine. But, we don't want to do worse than

Minimax Regret for Imprecise MDPs

We know that $r \in \mathcal{R}$, where $\mathcal{R}$ is the feassible reward set, which reflects the current knowledge of the reward.

$$R(f,r) = \max_{g \in \mathcal{F}} r.g - r.f MR(f,\mathcal{R}) = \max_{r \in \mathcal{R}} R(f,r) MMR(\mathcal{R}) = \min_{f \in \mathcal{F}} MR(f,\mathcal{R})$$

$R(f,r)$ is the regret of policy f (as represented by its visitation frequencies) relative to reward function r: it is simply the loss or difference in value between f and the optimal policy under r. $MR(f,R)$ is the maximum regret of f w.r.t. feasible reward set $\mathcal{R}$. Should we chose a policy with visitation frequencies f, $MR(f,R)$ represents the worst-case loss over all possible realizations of the reward function; i.e., the regret incurred in the presence of an adversary who chooses the r from $\mathcal{R}$ to maximize our loss. Finally, in the presence of such an adversary, we wish to minimize this max regret: $MMR(R)$ is the minimax regret of feasible reward set $\mathcal{R}$. This can be viewed as a game between a decision maker choosing f who wants to minimize loss relative to the optimal policy, and an adversary who chooses a reward to maximize this loss given the decision maker's choice of policy. Any f^* that minimizes max regret is a minimax optimal policy, while the r that maximizes its regret is the witness or adversarial reward function, and the optimal policy g for r is the witness or adversarial policy.

Regan and Boutilier, Regret based reward eliciting

Linear Programming Approach to Dynamic Programming

example

Assume an MDP with two states, and the reward/cost function of $\{1, 10\}$, and discount factor γ. By applying the *dynamic programming* operator (or equivalently, the Bellman operator) we have:

$$v = \{10 + \gamma v, 1 + \gamma v\} = Tv$$

$$J = \{10 + \gamma J, 1 + \gamma J\} = TJ$$

We are looking for the optimal policy. Minimization or maximization would work, or we can use the Bellman optimality equation.

$$v = \max_a \{10 + \gamma v, 1 + \gamma v\}$$

$$J = \min_a \{10 + \gamma J, 1 + \gamma J\}$$

In order to restructure this problem into the form of optimization problems, we can say we are trying to find the solution for these min/max problems

We are going explain why the max term turned into a min term, and vice versa. This is how it works for the case of v: Let's expand the equations:

$$\min v \quad s.t. \quad v \geq Tv$$

$$\max J s.t. \quad J \leq TJ$$

$$\min v \quad s.t. \quad v \geq Tv = \begin{cases} v \geq 10 + \gamma v \\ v \geq 1 + \gamma v \end{cases}$$

The constraint forces us to be more than $10 + \gamma v$, which is the maximum term in the function TV, then we have to pick exactely value $10 + \gamma v$ since we have a minimization. This is what we initially want our result to be. We already knew $10 + \gamma v$ is the term that maximizes our value function, and we were looking for a way around the problem to formulate it in an *constraint optimization* fashion, so that we can solve it later with **linear programming**.