Theory Analysis in Imitation Learning

Traditional Imitation Learning: Behavior Cloning

`Behavior Cloning` refers to an agent to mimic the expert's behavior in one task. However, there are few recover actions that refers that the expert correct his/her mistake, which means that the agent will continue to make mistakes when the agent firstly execute an action not belong to the training data. Therefore, this problem will cause to distribution shift and make a bad result.

Analysis

We train the policy under $p_{data}(O_t))$, and want to maxium $max_\theta E_{o_t \sim p_{data}(o_t)}[log\pi_\theta(a_t|o_t))]$.

But we test the policy under $P_{\pi_\theta}(O_t))$. And

$$P_{data}(O_t) \neq P_{\pi_\theta(O_t)}$$

We have several assumptions:

$$c(s_t, a_t) = \begin{cases} 0 & \text{if } a_t = \pi^*(s_t) \\ 1 & \text{otherwise} \end{cases}$$

that means you will get cost 1 if you make a mistake.

Our goal is to minimize:

$$E_{s_t \sim P_{\pi_\theta}(s_t)}[c(s_t, a_t))]$$

And we assum that the probability of agent makes a mistake is bound by ϵ :

$$\pi_\theta(a \neq \pi^*(s)|s) \leq \epsilon$$

for all $s \sim P_{train}(s)$.

We can show that:

$$E[\sum_t c(s_t, a_t)] \leq \epsilon T + (1-\epsilon)(\epsilon(T-1) + (1-\epsilon)(\dots))$$

So the expression $E \sim O(\epsilon T^2)$, that means the mistakes will increase quadratically and lead to bad results.

So look at the homework1:

Consider the problem of imitation learning within a discrete MDP with horizon T and an expert policy π^*. We gather expert demonstrations from π^* and fit an imitation policy π_θ to these trajectories so that

$$\mathbb{E}_{p_{\pi^*}(s)}\pi_\theta(a \neq \pi^*(s) \mid s) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{p_{\pi^*}(s_t)}\pi_\theta(a_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon,$$

i.e., the expected likelihood that the learned policy π_θ disagrees with the expert π^* within the training distribution p_{π^*} of states drawn from random expert trajectories is at most ε.

For convenience, the notation $p_\pi(s_t)$ indicates the state distribution under π at time step t while $p(s)$ indicates the state marginal of π across time steps, unless indicated otherwise.

1. Show that $\sum_{s_t} |p_{\pi_\theta}(s_t) - p_{\pi^*}(s_t)| \leq 2T\varepsilon$.

   Hint 1: in lecture, we showed a similar inequality under the stronger assumption $\pi_\theta(s_t \neq \pi^*(s_t) \mid s_t) \leq \varepsilon$ for every $s_t \in \text{supp}(p_{\pi^*})$. Try converting the inequality above into an expectation over p_{π^*}

   Hint 2: use the union bound inequality: for a set of events E_i, $\Pr[\bigcup_i E_i] \leq \sum_i \Pr[E_i]$

2. Consider the expected return of the learned policy π_θ for a state-dependent reward $r(s_t)$, where we assume the reward is bounded with $|r(s_t)| \leq R_{\max}$:

   $$J(\pi) = \sum_{t=1}^{T}\mathbb{E}_{p_\pi(s_t)}r(s_t).$$

   (a) Show that $J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T\varepsilon)$ when the reward only depends on the last state, i.e., $r(s_t) = 0$ for all $t < T$.

   (b) Show that $J(\pi^*) - J(\pi_\theta) = \mathcal{O}(T^2\varepsilon)$ for an arbitrary reward.

$$E_{P_{\pi^*(s)}}\pi_\theta(a \neq \pi^*(s)|s) \leq \epsilon$$

$$\sum_s P_{\pi^*(s)}\pi_\theta(a \neq \pi^*(s)|s) \leq \epsilon$$

the absolute difference between probability means that the union that the learned policy is different with expert policy. So use the hint 2:

$$|P_{\pi_\theta}(s_t) - P_{\pi^*}(s_t)| \leq 2\sum_{t=1}^{T}\sum_s P_{\pi^*(s)}\pi_\theta(a \neq \pi^*(s)|s) \leq 2T\epsilon$$

Look at the problem 2:

When the reward only depends on the last state:

$$J(\pi^*) - J(\pi_\theta) = E_{P_\pi(s_T))}r(s_T) - E_{P_{\pi_\theta}(s_T)}r(s_T)$$

$$J(\pi^*) - J(\pi_\theta) \leq |P_{\pi^*} - P_{\pi_\theta}|R_{max} \leq 2\epsilon T R_{max}$$

So, the $J(\pi^*) - J(\pi_\theta) = O(T\epsilon)$.

When the reward is an arbitary reward:

$$J(\pi^*) - J(\pi_\theta) \leq \sum_{t=1}^{T}|P_{\pi^*} - P_{\pi_\theta}|R_{max} \leq \sum_{t=1}^{T}2\epsilon R_{max} = O(\epsilon T^2)$$

When the reward function only depends on the final state and is zero for all other states, the difference in the expected return between the expert policy and the imitation policy is of the order $O(T\epsilon)$. This means that the difference in performance scales linearly with the horizon T and is directly proportional to the probabilityϵ of the imitation policy making a decision that differs from the expert policy. In practice, this suggests that if the imitation policy performs

slightly worse than the expert policy at each step, the total impact on the cumulative reward will be manageable over the horizon T because it's only the final state that matters.

For an arbitrary reward function, where rewards are accrued at every step, the difference in expected return scales with the square of the horizon $O(T^2\epsilon)$. This indicates that the impact of the imitation policy's deviations from the expert policy can compound at each step, leading to a quadratic increase in the total performance loss over time. This is a significant result, as it suggests that small mistakes can lead to a much larger cumulative discrepancy in the long run.

DAGRR

when we augment the data to make $P_{train} = P_{data}$, the problem caused by distribution shift will be alleviated.