Reinforcement Learning CS285 Notes and Homeworks

Andres Espinosa

December 17, 2024

Contents

1 Lecture 1

No notes taken during this lecture

2 Lecture 2

Supervised learning of behaviors and imitation learning

2.1 Part 1

2.1.1 Notation and Context

Given an observation o, our goal is to find a policy $\pi_\theta(a_t|o_t)$ that maps an observation to an action at a time t.

- $\mathbf{o}_t$ - observation
- $\mathbf{s}_t$ - state (note that this and observation tend to get confounded)
- $\mathbf{a}_t$ - action
- $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ - policy which can be deterministic or stochastic. We can generalize policies to be stochastic and then if we want a deterministic decision we can pick the most probable outcome.
- $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ - policy (fully observed)

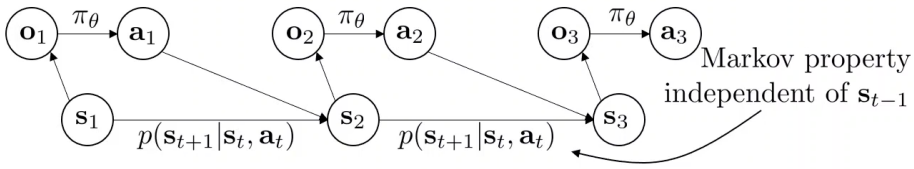

Figure 1: Graphical model showing the transition of observations, actions, and states

2.1.2 Imitation Learning

Imitation learning's goal is to get a policy from looking at observations and actions. Observation and action pairs $(\mathbf{o}_t, \mathbf{a}_t)$ are taken in as training data and fed into some supervised learning algorithm with the goal of outputting a policy $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$. Imitation learning, or behavior cloning, can be very flawed since it is difficult for it to adapt its knowledge to new situations. The supervised learning model will make a mistake that gets compounded as mistakes will cause it to deviate from the original trajectory. Being in a new trajectory increases the likelihood that the model will make more mistakes.
The issue of compounded mistakes is not present in traditional supervised learning since each observation in the data is IID. In a temporal sequence, like a control/reinforcement learning problem, the IID assumption does not hold. This issue can be avoided or mitigated by being smart about the way data is collected and augmented.

2.2 Part 2

2.2.1 Why does behavioral cloning fail?

If we have a policy $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ and a data set $(\mathbf{o}_t, \mathbf{a}_t)$. We can define the probabilities that we pick an action based off of an observation as $p_{data}(\mathbf{o}_t)$ and $p_{\pi_\theta}(\mathbf{o}_t)$. In this case, as we train using supervised

learning, we have the objective function

$$\max_{\theta} \mathbb{E}_{\mathbf{o}_t \ p_{data}(\mathbf{o}_t)}[\log \pi_\theta(\mathbf{a}_t|\mathbf{o}_t)] \tag{1}$$

Since there is a difference in the distribution that the model was trained on and the one that it will experience $p_{data}(\mathbf{o}_t)$ and $p_{\pi_\theta}(\mathbf{o}_t)$, this skews the output. A different objective function can be created using a cost function

$$c(\mathbf{s}_t, \mathbf{a}_t) = \begin{cases} 0 & if \mathbf{a}_t = \pi^*(\mathbf{s}_t) \\ 1 & otherwise \end{cases} \tag{2}$$

and we restructure our objective function to be

$$\text{minimize} \quad \mathbb{E}_{\mathbf{s}_t \ p_{\pi_\theta}(\mathbf{s}_t)}[c(\mathbf{s}_t, \mathbf{a}_t)] \tag{3}$$

We can analyze how effective behavioral cloning by identifying how likely you are at each time step t in a horizon T to make a mistake. If we assume that for any $\mathbf{s} \in D_{train}$, there is a probability $\mathbf{prob}[\pi_\theta(\mathbf{a} \neq \pi^*(\mathbf{s})|\mathbf{s})] \leq \epsilon$. Then if we want to find the expected value of our cost function across all time periods, we can find the below equation. Intuitively, the below equation works by first adding the probability we make a mistake each period to the probability we don't make a mistake on the first step and then every mistake every step after, etc.

$$\mathbb{E}[\sum_t c(\mathbf{s}_t, \mathbf{a}_t)] \leq \epsilon T + (1-\epsilon)(\epsilon(T-1)) \dots \tag{4}$$

This series increases at a rate $O(\epsilon T^2)$ in the worst case. This is a pessimistic worst case, so in order for behavioral cloning to work in practice, the cost functions and problem structure is usually built in a way to make the state easy to recover from.