CS 4/5789 18 Apr 2022 Prof Sarah Dean

Lecture 22: Imitation Learning

1) Motivation & Examples - Slides

2) Setting: Behavioural Cloning

Discounted Infinite Horizon MDP

$$\mathcal{M} = \{ \mathcal{S}, \mathcal{A}, P, r, \gamma \}$$

unknown!

possibly unobserved

Expert knows optimal policy π^* and we have a dataset

$$\mathcal{D} = (s_i^*, a_i^*)_{i=1}^{M} \sim d^{\pi^*}$$

Behavioral Cloning

Reduction to supervised machine learning.

Define some policy class Π

e.g. parametric policies $\Pi = \{ \pi_\theta \mid \theta \in \mathbb{R}^d \}$
where π_θ e.g. deep network with weights θ.

Then we estimate a policy with empirical risk minimization

$$\hat{\pi} = \underset{\pi \in \Pi}{\operatorname{argmin}} \sum_{i=1}^{M} \ell(\pi, s_i^*, a_i^*)$$

Depending on the problem, there are many choices of loss function.

- discrete action spaces: view problem as classification
- continuous action spaces: veiw problem as regression.

ex – Negative log likelihood

$$\ell(\pi, s, a) = -\log(\pi(a|s))$$

ex – Square loss

$$\ell(\pi, s, a) = \|\pi(s) - a\|_2^2$$

3) BC Analysis

Assumption: supervised learning is successful:

$$\mathbb{E}_{s \sim d_{\mu_0}^{\pi^*}} \left[\mathbb{1}\{\hat{\pi}(s) \neq \pi^*(s)\} \right] \leq \varepsilon$$

Notice: train and test distribution mismatch.
We train on data from distribution induced by π^*, and our error guarantee is with respect to this distribution.
However, our actual performance will be determined by data from the distribution induced by $\hat{\pi}$.

<u>Recall</u>: state distribution

$$d^{\pi}_{\mu_0}(s) = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h \mathbb{P}^{\pi}_h(s;\mu_0)$$

<u>Recall</u>: Performance Difference Lemma (PDL)

$$\mathop{\mathbb{E}}_{s\sim\mu}\left[V^{\pi}(s) - V^{\pi'}(s)\right] = \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi}_{\mu}}\left[\mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right]\right]$$

The difference in value is given by the advantage of π over π' averaged over the distribution induced by π.

<u>Theorem (BC Performance)</u>: Assume $r(s,a) \in [0,1]\ \forall\, s,a$.
And assume that supervised learning is ε-successful.
Then BC returns a policy $\hat{\pi}$ with

$$\mathop{\mathbb{E}}_{s\sim\mu}\left[V^{\pi^*}(s) - V^{\hat{\pi}}(s)\right] \leq \frac{2}{(1-\gamma)^2}\varepsilon$$

<u>Proof</u>:

$$\mathop{\mathbb{E}}_{s\sim\mu}\left[V^{\pi^*}(s) - V^{\hat{\pi}}(s)\right] = \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi^*}_{\mu}}\left[A^{\hat{\pi}}(s,\pi^*(s))\right] \quad \text{(PDL)}$$

optimal policy deterministic

(add 0)

$$= \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi^*}_{\mu}}\left(A^{\hat{\pi}}(s,\pi^*(s)) - A^{\hat{\pi}}(s,\hat{\pi}(s))\right)$$

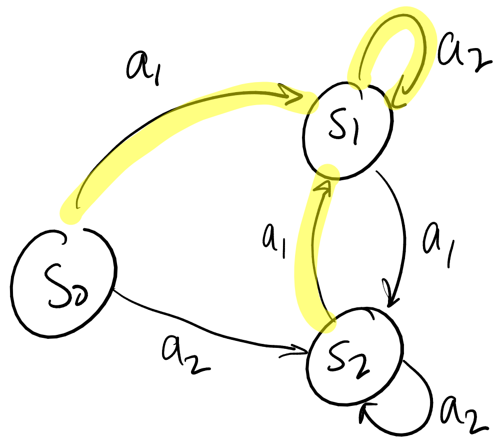
$\left(A \leq \frac{1}{1-\gamma}\right)$

$$\leq \frac{1}{1-\gamma}\mathop{\mathbb{E}}_{s\sim d^{\pi^*}_{\mu}}\left[\frac{2}{1-\gamma}\mathbb{1}\{\pi^*(s) \neq \hat{\pi}(s)\}\right]$$

(SL success)

$$\leq \frac{2}{(1-\gamma)^2}\varepsilon$$

Example: Distribution shift



s_0 initial state.

$$r(s,a) = \begin{cases} 1 & s = s_1 \\ 0 & \text{otherwise} \end{cases}$$

optimal policy is:

$$\pi^*(s) = \begin{cases} a_1 & s \neq s_1 \\ a_2 & s = s_1 \end{cases}$$

The distribution induced by the optimal policy is

$$d_{s_0}^{\pi^*}(s) = \begin{cases} 1-\gamma & s = s_0 \\ \gamma & s = s_1 \\ 0 & s = s_2 \end{cases}$$

Consider the following policy $\hat{\pi}$:

$$\hat{\pi}(s_0) = \begin{cases} a_1 & \text{w.p. } 1 - \frac{\varepsilon}{1-\gamma} \\ a_2 & \text{w.p. } \frac{\varepsilon}{1-\gamma} \end{cases} \qquad \pi(s_1) = a_2, \; \pi(s_2) = a_2$$

This policy could be retured by SL and has low error

$$\mathbb{E}_{s \sim d_{s_0}^{\pi^*}} \left[\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} \mathbb{1}\{a \neq \pi^*(s)\} \right] = \varepsilon$$

The quadratic error in performance:

$$V_{s_0}^{\pi^*} = \frac{\gamma}{1-\gamma} \qquad V_{s_0}^{\hat{\pi}} = \frac{\gamma}{1-\gamma} - \frac{\varepsilon\gamma}{1-\gamma} = \frac{\varepsilon\gamma}{(1-\gamma)^2}$$

is directly caused by our bad behavior in s_2 which prevents us from accumulating reward!