Lecture 23

Wednesday, November 11, 2020

CS59000 - RL

POMDP

Agenda

- Protocol
- Belief
- Hardness

Let's consider finite state, action, observation spaces. Then we can write P, O, P_1 in the following sense.

$P(x'|x,a)$

$O(y|x)$

$P_1(x)$

---

protocol:

At time step 1,
- $X_1 \sim P_1$
- $Y_1 \sim O(X_1)$
- choose action A_1
- succeed to $X_2 \sim P(X_1, A_1)$

$\vdots$

Wednesday, November 11, 2020

At time $t=1$, without any observation

P_1 : Prior

After observing Y_1, then what we think

about X_1? $\underbrace{b_1(x)}_{\text{belief at time } t=1} = P_r(X_1 = x \mid Y_1 = y_1)$

How to compute b_1?

$$b_1(x) = \frac{P_r(X_1 = x, Y_t = y_1)}{P_r(Y_1 = y_1)} = \frac{O(y_1 \mid x)\, P_1(X_1 = x)}{\sum_{x_1} O(y_1 \mid x_1)\, P_1(x_1)}$$

Note: $b_1(x) = P_r(X_1 = x \mid Y_1 = y_1, A_1 = a)$

If A_1 is chosen irrespect to x_1.

$$b_t(x) = P_r(X_t = x \mid \underbrace{y_1, a_1, \ldots, y_t}_{h_t})$$

$$= \frac{P_r(Y_t = y_t \mid X_t = x, h_{t-1}, a_{t-1})\, P_r(X_t = x \mid h_{t-1}, a_{t-1})}{P_r(Y_t \mid h_{t-1}, a_{t-1})}$$

Wednesday, November 11, 2020

$$b_t(x) = \frac{O(y_t|x) \sum_{x^i} P(X_t = x \mid \cancel{X}_{t-1} = x^i, a_{t-1}) \overbrace{Pr(X_{t-1} = x^i \mid h_{t-1}, a_{t-1})}^{b_{t-1}(x^i)}}{\sum_{x^i} \sum_{x^j} O(y_t \mid x^j) P(x^j \mid x^i, a_{t-1}) Pr(x^i \mid h_{t-1}, a_{t-1})}$$

$$\to Pr(X_{t-1} = x^i \mid h_{t-1}, a_{t-1}) = Pr(X_{t-1} = x^i \mid h_{t-1}) = b_{t-1}(x^i)$$

$b \in \Delta^X \Rightarrow$ we can have a policy as a function of b_t i.e. $A \sim \pi(b_t)$ $\Big\} \Rightarrow$ MDP on belief space

---

Optimal policy:

- Infinite horizon POMDP

  ↳ Undecidable Madani et al. 1999

- Fixed horizon:

given $y_1 \to b_1(x)$; $|Y|$ vectors.

$b_2(x)$; $|Y|^2 |A|$

$b_3(x)$; $|Y|^2 |A|^2$

Wednesday, November 11, 2020

$b_H(x)$; $|\mathcal{Y}|^H |\mathcal{A}|^{H-1}$

$\Rightarrow$ PSPACE- Complete

Papadimitriou and Tsitsiklis 1987

Memoryless policy.

NP- hard Vlassis et al. 2102

The optimal policy is stochastic Singh et al 1994

Memory- less $\Rightarrow$ Much better

Optimism in POMDP

$$\tilde{O}\left(D|\mathcal{X}|^{3/2}\sqrt{|\mathcal{A}||\mathcal{Y}|T}\right)$$

$\rightarrow$ Azizzadenesheli 2016

So many open Problem.

off-policy learning.

π_b : behavioural policy ;

- using data generated by π_b can we say how well a new π does
- Can we find an optimal policy.

Remember the Contextual bandit, i.e. a Fixed horizon MDP with horizon 1.

Consider a stationary Contextual bandit. with context set $\mathcal{X}$, action set $\mathcal{A}$, and a measure μ on $\mathcal{X}$. with R, the reward kernel and mean $\bar{r}$; $0 \leq \bar{r} \leq 1$

$0 \leq r \leq 1$

Protocol:

- Draw $X \sim \mu$
- Draw $A \sim \pi(X_t)$
- recive $r \sim R(X_t, A_t)$

For π ; $\eta(\pi) = E_\mu\Big[E_\pi[\,r \mid X\,]\Big]$

$$= E_\mu\Big[E_\pi[\,\bar{r}(X, A) \mid X\,]\Big]$$

After following π_b for T time step,

we have $D_T = \{ X_t, A_t, r_t \}_{t=1}^{T}$

what is $\eta(\pi_b)$?

$$\hat{\eta}(\pi_b) = \frac{1}{T} \sum r_t \quad \Bigg| \quad |\hat{\eta}(\pi_b) - \eta(\pi_b)| \leq \sqrt{\frac{1}{2T} \log \frac{2}{\delta}}$$

with probability at least $1-\delta$

How about $\eta(\pi)$?

For simplicity, Let $\mathcal{A}$ to be a finite set. space and $\pi(dA; x) = \pi(A; x)\, dA$

Define $\hat{r}_t(X_t, a) = r_t \dfrac{1(A_t = a)}{\pi_b(A_t; X_t)}$

↳ the inverse propensity score (IPS)

$$- E_{\pi_b}[\hat{r}_t(X_t, a) \mid X_t] = \int_{\mathcal{A}} \frac{\bar{r}(X_t, A_t)\, 1(A_t = a)}{\cancel{\pi_b(A_t; X_t)}} \cancel{\pi_b(A_t; X_t)}\, dA_t$$

$$= \bar{r}(X_t, a)$$

Wednesday, November 11, 2020

now we can use $\hat{r}$ to estimate $\eta(\pi)$

$$\underline{\eta_{IPS}}(\pi) = \frac{1}{T}\sum_{t=1}^{T} E_\pi[\hat{r} \mid X_t]$$

$$= \frac{1}{T}\sum_{t=1}^{T} \int \hat{r}(X_t, A)\,\pi(A; X_t)\,dA$$

$$= \frac{1}{T}\sum_{t=1}^{T} r_t \frac{\pi(A_t; X_t)}{\pi_b(A_t; X_t)}$$

why it is good?

$$\eta(\pi) = E_\mu\left[E_\pi[r/X]\right] = E_\mu E_{\pi_b}\left[\underbrace{\frac{\pi}{\pi_b}}_{\text{Importance weight}} r/X\right]$$

justification,

$$\eta(\pi) = E_\mu\left[E_\pi\left[\bar{r}(X, A) \mid X\right]\right]$$

$$= E_\mu\left[\int \bar{r}(X, A)\,\pi(A; X)\,dA\right]$$

$$= E_\mu\left[\int \bar{r}(X, A)\frac{\pi(A; X)}{\pi_b(A; X)}\,\pi_b(A; X)\,dA\right]$$

$$= E_\mu\left[E_{\pi_b}\left[\bar{r}(X, A)\frac{\pi(A; X)}{\pi_b(A; X)} \mid X\right]\right]$$

Wednesday, November 11, 2020

$\Rightarrow$ therefore, the empirical estimate

$$\hat{\eta}(\pi) = \sum_{t=1}^{T} r(X_t, A_t) \frac{\pi(A_t; X_t)}{\pi_b(A_t; X_t)}$$

i.e. an Monte Carlo estimate of the integral

$$\Rightarrow \hat{\eta}_{IPS}(\pi) = \hat{\eta}(\pi)$$

How good is this estimation?

Assume $\omega_{max} := \text{ess sup} \frac{\pi}{\pi_b} \leftarrow$

$$|\hat{\eta}(\pi) - \eta(\pi)| \leq \omega_{max} \sqrt{\frac{1}{2T} \log \frac{2}{\delta}}$$

w.p. at least $1-\delta$.