Monday, November 16, 2020

Lecture 24

CS_59000-RL

Contexual Bandit & MDP

Agenda

- off policy policy evaluation
- off policy policy optimization

From last lecture: $D := \{X_i, A_i, r_i\}_{i=1}^{T}$ generated by π_b

Estimate the performance of π

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

$$\omega_{max} := \operatorname{ess\,sup} \frac{\pi}{\pi_b}$$

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

with probability at least $1-\delta$. (From Hoeffding)

Monday, November 16, 2020

what is the variance of $\hat{\eta}(\pi)$?

let $\sigma^2(x,a)$ denote $Var[r(X,A) \mid X, A](x,a)$

$$\Rightarrow \underset{\mu,\pi_b}{Var}[\hat{\eta}(\pi)] = \frac{1}{T} \underset{\mu,\pi_b}{Var}[\omega(X,A) r]$$

$\omega(X,A) \to$ Importance weight function; i.e. $\frac{\pi(X,A)}{\pi_b(X,A)}$

Using the law of total variance

$$\Rightarrow Var[X_1] \underset{a.s.}{=} E[Var[\underline{X_1} \mid X_2]] + Var(E[\underline{X_1} \mid X_2])$$

$$\Rightarrow \frac{1}{2} \underset{\mu,\pi_b}{Var}[\omega(X,A) r]$$

$$= \frac{1}{T} \underset{\mu,\pi_b}{E}[Var[\underline{r}\ \underline{\omega(X,A)} \mid \underline{X,A}]]$$

$$+ \frac{1}{T} \underset{\mu,\pi_b}{Var}(E[\omega(X,A) r \mid X,A])$$

$$= \frac{1}{T} \underline{E_{\mu,\pi_b}}[\underline{\omega^2(X,A)} \sigma^2(X,A)] + \frac{1}{T} \underset{\mu,\pi_b}{Var}[\omega \bar{r}]$$

Monday, November 16, 2020

$\Rightarrow$ One can use Bernstine or empirical Berstine to come up with maybe a better bound

---

Lets imagine some one gives us $\underbrace{r'(x,a)}$ $\forall x, y \in \mathcal{X} \times A$ as a surrogate. Imagine $r' \simeq \bar{r}$

means, somewhat model based.

Doubly robust estimator

$$\hat{r}_t^{DR}(X_t, a) = \underline{r'(X_t, a)} + \underline{\left(r_t - r'(X_t, a)\right)} \frac{1(A_t = a)}{\pi_b(A_t; X_t)}$$

(finite action spaces)

HW: Similar to $\hat{r}_t(X_t, a)$, show that $\hat{r}_t^{DR}(X_t, a)$ is also an unbiased estimator of $\bar{r}(X_t, a)$

$$\Rightarrow Var^{DR}(\pi) = \frac{1}{T} E_{\mu, \pi_b}\left[\sigma^2(X, A)\, \omega^2(X, A)\right]$$

$$+ \frac{1}{T} E_\mu\left[Var_{\pi_b}\left[\underline{\omega(X, A)}\left(\underline{\bar{r}(X, A) - r'(X, A)}\right) \middle| X\right]\right]$$

$$+ \frac{1}{T} Var_\mu\left[E_{\pi_b}\left(\bar{r}(X, A)\, \omega(X, A) \,|\, X\right)\right]$$

Some one gives us r'

we can learn it ourselve:

E.g. given $D' \Rightarrow \inf_{f \in F} \frac{1}{T'} \sum \left(f(X_t, A_t) - r(X_t, A_t) \right)^2$

↳ a class of functions.

$\Rightarrow$ Set r' to be an $f \in F$ with low loss

Sometime model based + IS might end up with better result.

---

policy optimization.

- Give a set of policy Π_N with N policies
- which one is the best?

we are looking for $\max_{\pi \in \Pi_N} \eta(\pi) \rightarrow \pi^*$

E.g. we use IS. $\Rightarrow \forall \pi \in \Pi_N \rightarrow \underline{\hat{\eta}(\pi)}$

$\max_{\pi \in \Pi_N} \hat{\eta}(\pi) \rightarrow \hat{\pi}^*$

Monday, November 16, 2020

How good is $\hat{\pi}^* \Rightarrow \eta(\hat{\pi}^*) \Leftrightarrow \eta(\pi^*)$

Theorem

$$\eta(\hat{\pi}^*) \geq \eta(\pi^*) - \omega_{max}\sqrt{\frac{2}{T}\log\frac{2N}{\delta}}$$

proof: Using union bound and Hoeffding's

$$\rightarrow \max_{\pi\in\Pi_N} \left|\hat{\eta}(\pi) - \eta(\pi)\right| \leq \omega_{max}\sqrt{\frac{1}{2T}\log\frac{2N}{\delta}}$$

with prob at least $1-\delta$.

where $\omega_{max} = \max_{\pi\in\Pi_N} \operatorname{esssup}\frac{\pi}{\pi_b}$

$$\rightarrow \eta(\hat{\pi}^*) \geq \underbrace{\hat{\eta}(\hat{\pi}^*)} - \omega_{max}\sqrt{\frac{1}{2T}\log\left(\frac{2N}{\delta}\right)}$$

$$\geq \hat{\eta}(\pi^*) - \omega_{max}\sqrt{\frac{1}{2T}\log\left(\frac{2N}{\delta}\right)}$$

$$\geq \eta(\pi^*) - \omega_{max}\sqrt{\frac{2}{T}\log\left(\frac{2N}{\delta}\right)}$$

Bingo

Monday, November 16, 2020

Now for MDPs.

Consider a fixed horizon MDP

$$M(\mathcal{X}, A, P, R, P_1, \gamma, H) ; \gamma \in [0,1]$$

Let $\rho_h = (X_1, A_1, r_1, \ldots, X_h, A_h, r_h)$ in an episode.

$\rightarrow$ stage h

$$\rightarrow \eta(\pi) = E_{P_\pi}\left[\sum_{h=1}^{H} \gamma^{h-1} r_h\right]$$

we are given data from $\pi_b \Rightarrow \eta(\pi)$

If $P_\pi \ll P_{\pi_b} \Rightarrow \eta(\pi) = E_{P_{\pi_b}}\left[\frac{dP_\pi}{dP_{\pi_b}}\left(\sum_{h=1}^{H} \gamma^{h-1} r_h\right)\right]$

E.g. For special case of Euclidean space:

$$dP_\pi = P_1(X_1)\,\pi(A_1; X_1)\,P(X_2; X_1, A_1) \ldots dX_1 dA_1 \ldots$$

$$\hookrightarrow \frac{dP_\pi}{dP_{\pi_b}} = \prod_{h=1}^{H} \frac{\pi(A_h; x_h)}{\pi_b(A; x_h)} \quad ; \rho_h = \frac{\pi(A_h; X_h)}{\pi_b(A_h; X_h)}$$

Monday, November 16, 2020

Simply for step-IS we have ↓

$$\eta(\pi) = E_{\rho_{\pi_b}}\left[\sum_{h=1}^{H} \gamma^{h-1} \Big(\underbrace{\prod_{h'=1}^{h} \rho_{h'}}\Big) r_h\right]$$

$$V_H = \bar{r} \rightarrow V_{H-1} = r + P\, V_H$$

$$V_{H-2} = r + P\, V_{H-1}$$

We are given data π_b : $D := \{(X_h^t, A_h^t, r_h^t)\}_{h=1, t=1}^{H, T}$

$$\hat{V}_h(x) = \frac{1}{T_h(x)} \sum_{t=1}^{T} \rho_h^t \left(r_h^t + \gamma \hat{V}_{h+1}\right) I(X_h^t = x)$$

$$\hookrightarrow \hat{V}_H(x) = \frac{1}{\underbrace{T_H(x)}} \sum_{t=1}^{T} \underbrace{\rho_H^t(X_H^t, A_H^t)}\ \underbrace{r_H^t}\ I(X_H^t = x)$$

$$\hat{V}_{H-1}(x) = \frac{1}{T_{H-1}(x)} \sum_{t=1}^{T} \rho_{H-1}^t (X_{H-1}^t, A_{H-1}^t)$$
$$\left(r_{H-1}^t + \gamma \hat{V}_H(X_H^t)\right) I(X_{H-1}^t = x)$$

Monday, November 16, 2020

DR:

$$\hat{V}_h^{DR}(x_t) = \underbrace{V_h'(x_t)} + \frac{1}{T_h(x_t)} \sum_{t=1}^{T} \rho_h^t \left(r_h^t + \gamma \hat{V}_{h+1}^{DR} - Q'(x_t, A_t) \right) I(X_h^t = x)$$

Both of these estimators are unbiased.

What are the variances:

$$\mathrm{Var}\left[\hat{V}_h^{DR}\right] = \mathrm{Var}\left[V_h'(X_h)\right]$$

$$+ E\left[\mathrm{Var}\left[\rho_h \left(Q'(X_h, A_h) - Q(X_h, A_h)\right) \middle| X_h\right]\right]$$

$$+ E\left[\rho_h^2 \, \mathrm{Var}\left[r_h \mid X_h, A_h\right]\right]$$

$$+ E\left[\gamma^2 \rho_h^2 \, \mathrm{Var}\left[V_{h+1}^{DR} \mid X_h, A_h\right]\right]$$

Proof: Final exam

Monday, November 16, 2020

Model based (UCRL2)

Model free → Value based (Q-learning, SARSA)
→ Policy based

Model based

↳ $M \Rightarrow Q \longrightarrow \pi \Rightarrow$ good $\eta(\pi)$

Model free - Value based

$\rightarrow Q \Rightarrow \pi \Rightarrow$ good $\eta(\pi)$

Model free - policy based

$\pi \Rightarrow$ good $\eta(\pi)$

$$\max_{\pi \in \Pi} \eta(\pi)$$

⇒ Policy gradient

$$\underbrace{\nabla_\pi \eta(\pi)\Big|_{\pi=\pi_t}} \Rightarrow \pi_{t+1} \leftarrow \mathrm{Proj}\left(\pi_t + \alpha \nabla_\pi \eta(\pi)\Big|_{\pi=\pi_t}\right)$$