IEOR 8100-001: Learning and Optimization for Sequential Decision Making 02/22/16

Lecture 10: Contextual Bandits

Guest Lecture by Alekh Agarwal Scribed by: Vashist Avadhanula

In this lecture, we consider a setting that generalizes linear bandits. Specifically, we will study contextual bandits, where in each time period, we observe a context (features), take an action and observe reward that is dependent on the context as well as our action. The objective is to design an optimal policy that would suggest an action based on the observed context.

1 Contextual Bandits

Consider T time periods, at every time period t,

- Observe a context $x_t \in R^d$.
- Take action $a_t \in \{1, \cdots, N\}$.
- Observe reward $r_t(a_t) \in [0, 1]$.

Let $\mathcal{X}$ be the set of valid of contexts and let $\mathcal{A} = \{1, \cdots, N\}$ be the set of possible actions. The objective is to come up with a policy $\pi : \mathcal{X} \to \mathcal{A}$ that assigns an action to every valid context. Let Π denote the set of all valid policies. The objective is to minimize the regret, which is defined as

$$R_T = \left(\max_{\pi \in \Pi} \sum_{t=1}^{T} r_t(\pi(x_t)) \right) - \sum_{t=1}^{T} r_t(a_t)$$

Observe that when there is no context, this setting is very similar to the Multi-armed Bandit problem. We can think context in terms of partial feedback, for example consider the Netflix case we considered in the Linear Bandits lecture. In the Netflix example, context would be user profile information that would be helpful in suggesting movies. As an example for Π, consider a setting, where every action a, has a weight vector w_a associated with it and the policy is to play the action that maximizes the inner product $x_t \cdot w_a$. In this setting, the Π denotes the collection of all feasible weight vectors.

Naively extending the MAB approach to this problem will lead us to a bound $O(\sqrt{|\Pi| \log \Pi})$, which is not desirable since the value of $|\Pi|$ can take a large value. We will now extend the Exp-3 approach to achieve a better regret bound.

1.1 Exp4:

Recall that in Exp3 approach, we played action a with probability $P_t(a) \propto \exp\left(\eta \hat{R}_t(a)\right)$, where

$$\hat{R}_t(a) = \sum_{s=1}^{t} \frac{r_s(a)}{P_s(a)} \mathbb{1}(a = a_s).$$

We will take a similar approach to exp3, where we play policy π with probability $P_t(\pi) \propto \exp\left(\eta \hat{R}_t(\pi)\right)$, where $\hat{R}_t(\pi)$ is given by

$$\hat{R}_t(\pi) = \sum_{s=1}^{t} \frac{r_s(\pi(x_s))}{P_s(\pi(x_s))} \mathbb{1}\left\{a_s = \pi(x_s))\right\},$$

and $P_s(\pi(x_s)) = \sum_{\pi':\pi'(x(s))=\pi(x(s))} P_s(\pi')$.

We will use the following more general bound from [**?**] (for EXP3) to bound the regret of Exp-4. If we play policy π with probability $P(\pi)$, then

$$R_T(P) \leq \eta \sum_{t=1}^{T} P_t \cdot \hat{R}_t^2 + \frac{1}{\eta} KL\left(P \| U\right),$$

where U is the uniform distribution over all policies. From the above inequality, we have

$$\begin{aligned}
E(R_T(P) &\leq \eta \sum_{t=1}^{T} \sum_{\pi} E(P_t(\pi). \cdot \hat{R}_t^2(\pi))) + \frac{1}{\eta} KL(P\|U) \\
&= \eta \sum_{t=1}^{T} \sum_{a=1}^{N} \sum_{\pi:\pi(x_t)=a} \frac{r_t^2(a) P_t(\pi(x_t))}{P_t^2(\pi(x_t))} + \frac{1}{\eta} KL(P\|U) \\
&= \eta \sum_{t=1}^{T} \sum_{a=1}^{N} r_t^2(a) + \frac{1}{\eta} KL(P\|U) \\
&\leq \eta NT + \frac{\log|\Pi|}{\eta}
\end{aligned}$$

Optimizing the above bound over η, we have $E(R_T(P)) \leq O(\sqrt{NT \log|\Pi|})$. This bound is tight in the sense we can obtain a matching lower bound, but this policy is not implementable in practice, since the number of feasible policies can be large and it may not be computationally feasible to normalize their probabilities. So, now we consider a more structured setting.

2 Stochastic Rewards

Here, we assume that the contexts x_t and rewards r_t come from an unknown distribution, i.e. for each t, $(x_t, r_t) \sim \mathcal{D}$. Here the objective is to minimize the regret and this problem is also known in literature as ERM, empirical risk minimization or empirical reward maximization.

We first consider the setting when the values of $r_t(a)$ are known for $t = 1, \cdots, n$, then the greedy policy

$$\hat{\pi}_{greedy} = \underset{\pi \in \Pi}{argmax} \sum_{t=1}^{n} r_t(\pi(x_t)),$$

regret can be bounded by using chernoff and union bounds. Specifically, we can show that with at least probability $1 - \delta$

$$E\left(r(\hat{\pi}_{greedy})\right) \geq \max_{\pi \in \Pi} E\left[r(\pi)\right] - O\left(\sqrt{\frac{\log|\Pi|}{n\delta}}\right)$$

No we extend the ideas of importance sampling to the setting when $r_t(a)$ is not known. We initially explore to better understand the rewards and then will use the greed policy to exploit.

- For $t = 1, \cdots, n$, Pick $a_t \in \{1, \cdots, n\}$ uniformly. We observe $r_t(a_t)$.
- For $t = n+1, \cdots, T$, we play the greedy policy

$$\hat{\pi}_{greedy} = \underset{\pi \in \Pi}{argmax} \sum_{t=1}^{n} \frac{r_t(\pi(x_t))}{P_t(\pi(x_t))} \cdot \mathbb{1}(\pi(x_t) = a_t).$$

Let $\hat{r}_t(a) = \frac{r_t(\pi(x_t))}{P_t(\pi(x_t))} \cdot \mathbb{1}(\pi(x_t) = a)$. We have by importance sampling, $E(\hat{r}_t(a)) = r_t(a)$ and by Hoeffding inequality,

$$\forall\, \pi \in \Pi \left| \sum_{t=1}^{n} \hat{r}_t(\pi(x_t)) - E(r(\pi(x))) \right| \leq N\sqrt{\frac{1}{2n} \log \frac{2|\Pi|}{\delta}} \text{ w.p. } \geq 1 - \delta$$

$$\implies E(r(\hat{\pi})) \geq \max_{\pi \in \Pi} E(r(\pi(x))) - N\sqrt{\frac{1}{2n} \log \frac{2|\Pi|}{\delta}}$$

Therefore, we have $Regret \leq n + (T - n)\frac{N}{\sqrt{n}}\sqrt{\log \frac{2|\Pi|}{\delta}}$.

References

[1] The Nonstochastic Multiarmed Bandit Problem, Peter Auer, Nicol Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. SIAM Journal on Computing, Volume 32 Issue 1, 2003 Pages 48 - 77.

[2] Y. Freund, R.E.Schapire 1996. *A Decision-Theoretic Generalization of On-Line Learning and Application to Boosting.* Journal of Computer and System Sciences.