CS229T/STATS231: Statistical Learning Theory

Lecturer: Tengyu Ma
Scribe: Eric Wang (ejwang)

Lecture # 10
December 3, 2018

1 Review

Last time, we talked about the UCB algorithm and the constant multi-armed bandit problem. In this lecture we'll give the regret bound analysis for that problem and talk about the Bayesian regret setting.

Last time, we defined the *lower confidence bound*

$$LCB_t * (a) = \hat{\mu}_{t-1}(a) = 2\sqrt{\frac{\log T}{n_{t-1}a}}.$$

We choose the action a_t according to the LCB:

$$a_t = \arg\min_{a \in [n]} LCB_t(a).$$

We call this principle "optimism in the face of uncertainty"; we're looking at the most optimistic outcome consistent with the existing data. We claimed that this is a reasonable confidence bound: *for any action a with probability $> 1 - \frac{1}{T}$, for all t in $\{1, \ldots, T\}$ we have*

$$|\hat{\mu}_t(a) - \mu(a)| \leq 2\sqrt{\frac{\log T}{n_t(a)}},$$

$$\mu(a) = \left[\hat{\mu}(a) - 2\sqrt{\frac{\log T}{n_t(a)}}, \hat{\mu}_t(a) + 2\sqrt{\frac{\log T}{n_t(a)}}\right].$$

This can be viewed as a confidence interval that scales as $1/\sqrt{n}$. Finally, we had this theorem:

Theorem 1. *The regret of the UCB algorithm is bounded above by*

$$\sum_{a:\Delta a > 0} O\left(\frac{\log T}{\Delta a}\right).$$

(In comparison, the explore-then-exploit algorithm has regret $\sum O\left(\frac{\log T \cdot \Delta a}{\Delta^2}\right)$).

Recall that $\Delta a = \mu(a) - \mu(a^*)$, $\Delta = \min_{a:\Delta a>0} \Delta a$. Hence $\sum O\left(\frac{\log T \cdot \Delta a}{\Delta^2}\right) \geq \sum O\left(\frac{\log T}{\Delta a}\right)$. The proof is pretty straighforward given the earlier claim:

Proof. Recall that the regret is equal to

$$\sum [n_T(a)] \cdot \Delta a.$$

Thus it suffices to show that $\mathbb{E}\left[n_T(a)\right] \geq O(\frac{\log T}{\delta a^2})$ for all a. Fix one such a. We define T_0 analogously in the explore-and-exploit algorithm:

$$T_0 \triangleq \frac{20 \log T}{\Delta a^2}; \mathbb{E}\left[n_T(a)\right] \leq T_0 + \sum_{t=T_0}^{T} \mathbb{E}\left[\mathbf{1}(a_t = a, n_{t-1}(a) \geq T_0)\right].$$

This is because

$$\begin{aligned}
\mathbb{E}\left[n_T(a)\right] &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{1}(a_t = a)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbf{1}(a_t = a, n_{t-1}(a) < T_0)\right] + \sum_{t=1}^{T} E\left[\mathbf{1}(a_t = a, n_{t-1}(a) \geq T_0)\right]
\end{aligned}$$

Suppose the events E_a, E_{a^*} in the claim happen. Then

$$\begin{aligned}
LCB_t(a^*) &\leq \mu(a^*) \\
LCB_t(a) &= \hat{\mu}_{t-1}(a) - 2\sqrt{\frac{\log T}{n_{t-1}(a)}} \\
&\geq \mu(a) - 2\sqrt{\frac{\log T}{n_{t-1}(a)}} - 2\sqrt{\frac{\log T}{n_{t-1}(a)}} \\
&\geq \mu(a^*) + \Delta a - 4\sqrt{\frac{\log T}{n_{t-1}(a)}} \\
&\geq \mu(a^*) + \Delta a - 4\sqrt{\frac{\log T}{T_0}} \\
&> \mu(a^*)
\end{aligned}$$

Hence $LCB_t(a) \geq LCB_t(a^*)$ and $a_t \neq a$. Therefore

$$\begin{aligned}
\mathbb{E}\left[n_T(a)\right] &\leq T_0 + \sum_{t=1}^{T} \Pr\left[\overline{E_a \wedge E_{a^*}}\right] \\
&\leq T_0 + \frac{2}{T}(T - T_0) \\
&\leq T_0 + 2 \leq 2T_0 \leq O\left(\frac{\log T}{\Delta a^2}\right).
\end{aligned}$$

□

Last time, we provided an informal proof of the claim using Hoeffding's inequality. However, the conditions of Hoeffding's inequality weren't actually satisfied. The number of random variables was itself a random variable, for instance. We can get around that by using a high-probability bound:

2 Rigorizing the proof from last time

Consider the following provess:

- Generate $Z_1, \ldots, Z_T \sim D_a$ in advance, before the game starts.
- Every time action a is taken, return the next unused Z_j as a loss.

$$\ell_t(a) \triangleq Z_{n_{t-1}(a)+1}.$$

Although the process of generation is very complicated, the process of generating the Zs is still independent. Now, by Hoeffding's inequality and the union bound,

$$\forall j = 1, \ldots, T : \left| \frac{1}{j} \sum_{k=1}^{j} Z_k - \mu(a) \right| \le 2\sqrt{\frac{\log T}{j}}.$$

Hence

$$\hat{\mu}_t(a) - \frac{1}{n_t(a)} \sum_{i=1}^{T} \ell_i(a) \mathbf{1}(a_i = a) = \frac{1}{n_t(a)} \sum_{j=1}^{n_t(a)} Z_j$$

and $|\hat{\mu}_t(a) - \mu(a)| \le 2\sqrt{\frac{\log T}{n_t(a)}}$, completing the proof.

3 Bayesian Multi-Armed Bandit Problem

The general stochastic bandit problem: We have:

- A model parameter $\theta \in \Theta$. In the original problem, this was $\mu(1), \ldots, \mu(N)$.
- An action a in a family $\mathcal{A}$. In the original problem, this was $a \in [N]$.
- A distribution of loss, $D(a, \theta)$. In the original problem, this was D_a.

There is a ground truth parameter, θ^*. Now, at time t, if action a_t is chosen, we observe a loss $L_{a_t,\theta^*} \sim D(a, \theta^*)$. The *optimal action* is a function

$$a^* : \Theta \to \mathcal{A} \text{ such that}$$
$$a^*(\theta) = \arg\min_{a \in \mathcal{A}} \mathop{\mathbb{E}}_{L \sim D(a,\theta^*)} [L].$$

For simplicity, we work with unique optimal actions. In the multi-armed bandit problem, this was equal to $\arg\min_{a \in [n]} \mu(a)$. For a ground truth θ^* and actions $a_1, \ldots, a_T$, we define

$$\text{Regret}(\theta^*, a, \ldots, a_T) = \mathop{\mathbb{E}}_{L_{a_t,\theta^*} \sim D(a_t,\theta^*); L_{a_t^*,\theta^*} \sim D(a^*(\theta^*),\theta^*}\left[\sum_t L_{a_t,\theta^*} - \sum_t L_{a^*(\theta^*),\theta^*} \right].$$

We now turn to the Bayesian setting, where θ^* is posited to be drawn from some distribution Q, the *prior of the model parameter*. Let $A_1, \ldots, A_T$ be the actions taken by the algorithm. Then the Bayesian regret of the algorithm is defined as follows:

$$\text{Regret} = \mathop{\mathbb{E}}_{\theta^* \sim Q} \left[\mathop{\mathbb{E}}_{A_1,\ldots,A_T} [\text{Regret}(\theta^*, A_1, \ldots, A_T)] \right] \quad = \mathop{\mathbb{E}}_{\cdots} \left[\sum_t L_{a_t,\theta^*} - \sum_t L_{a^*(\theta^*),\theta^*} \right]$$

There's nothing special about this formulation. But this is a reasonable setting, and it exhibits some fairly nice behavior.

4 Solving the Bayesian MAB problem

One algorithm is *Thompson sampling*. "Repeatedly updating the posterior, drawing ground truth from the posterior, then playing the best action according to that truth."

Let $\mathcal{F}_{t-1}$ as shorthand for the random variables observed so far:

$$\mathcal{F}_{t-1} = \{A_1, L_1, \dots, A_{t-1}, L_{t-1}\}.$$

On each iteration,

- Compute the distribution

$$p_r(\theta) = \Pr(\theta^* = \theta | \mathcal{F}_{t-1}).$$

  This is the posterior of $\theta^* | \mathcal{F}_{t-1}$.

- Sample $\theta_t \sim p_t$.
- Play $a^*(\theta_t)$.

Next time, we'll bound the regret of Thompson sampling, as determined in a paper by [Russo and van Roy, '16].

5 Info Theory Background

(We need this to even state our bound.) Let $\mathcal{X}$ be a finite set and let X be a random variable over $\mathcal{X}$. The *entropy* of X is a measure of the amount of uncertainty, and it is given by

$$H(X) = -\sum_{x \in \mathcal{X}} Pr(X = x) \log \Pr(X = x).$$

It is a fact that $0 \leq H(X) \leq \log |X|$ and this follows from the concavity of the logarithm.

Let X, Y be random variables, The *conditional entropy* $H(X|Y)$ is defined as

$$\sum_y H(X|Y = y) \Pr(Y = y).$$

We define the *mutual information* or "entropy reduction" between X and Y is

$$I(X;Y) \triangleq H(X) - H(X|Y).$$

"How much entropy have I lost by observing Y?"

5.1 Properties of conditional information and mutual information

- $H(X|Y) = H((X, Y)) - H(Y)$.
- $I(X;Y) = I(Y;X) = H(X) + H(Y) - H(X, Y)$
- $I(X;Y) \geq 0 \Leftrightarrow H(X|Y) \leq H(x)$
- $I(X;Y) \geq H(X)$ (because $H(X|Y) \geq 0$).
- $I(X;Y) = 0 \Leftrightarrow X, Y$ independent.