A simple Research on Multi-Armed Bandit

Wang Pengyuan
Xi'an, Shaanxi, Northwestern Polytechnical University

E-mail: wpy3458@foxmail.com

This document presents a number of hints about how to set up your *Science* paper in LaTeX . We provide a template file, `scifile.tex`, that you can use to set up the LaTeX source for your article. An example of the style is the special {`sciabstract`} environment used to set up the abstract you see here.

1 The Advantages and Disadvantages of each Compared to the Different Multi-Armed Bandit Methods

In the section, I used *average reward*, *optimal action*, *percentage of top 3 superior action*, *reward variance* and so on to analyze the different Multi-Armed Bandit Methods. There are two main parts: methods, experiments and analysis.

1.1 Methods

There I attempted the **Value Estimation**(including *Greedy*, ϵ*-Greedy*, *Optimistic Initial Value*, UCB methods), **Preference Estimation**(including *Gradient method*), **Bayesian Estimation**(including Thompson Sampling method). The following are the brief introduction for their characters.

Greedy For every state, the agent just select the action whose reward is highest.

ϵ-Greedy The method is an improvement on *Greedy*. The only different is that the action will be chosen at random by the agent with ϵ probability.

Optimistic Initial Value Namely the estimate value is initialized higher than the real value.

UCB The choose policy is different with others. The agent choose action based on $A_t \doteq \underset{a}{argmax}\left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)})}\right]$, which is proved in Appendix A.

Gradient method It is based on the idea of gradient ascent and uses a preference function $H_t(a)$ to select actions. The proof and understanding are in Appendix B.

Thompson Sampling method Update q values using posterior probabilities based on Bayesian theory.

1.2 Experiments

1.3 Analysis

Appendix A

Q: Why UCB formula is $A_t \doteq \underset{a}{argmax}\left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}})\right]$? Why not $A_t \doteq \underset{a}{argmax}\left[Q_t(a) + c\frac{e^t - e^{-t}}{e^t + e^{-t}}\right]$? Why not others?

A: In real life, we can not get the exact value of every action. Namely $\tilde{q} \approx q$, where $\tilde{q}$ is the value we estimated and q is real value. There we can build the model $\tilde{q} - \triangle \leq q \leq \tilde{q} + \triangle$ *(1)*. According the model, we are optimistic that each action can be rewarded with $\tilde{q} + \triangle$, which is called UCB. So we only need to find $\triangle$ to represent the UCB.

There we need *Chernoff-Hoeffding Bound*

theorem 1 (Chernoff-Hoeffding Bound) $P\left\{|\tilde{p} - p| \leq \delta\right\} \geq 1 - 2e^{-2n\delta^2}$

When δ get the value $\sqrt{2\ln t/n}$, we can get

$$P\left\{|\tilde{p} - p| \leq \sqrt{2\ln t/n}\right\} \geq 1 - \frac{2}{T^4} \tag{1}$$

Therefore, we can get the formula $\tilde{p} - \sqrt{2\ln t/n} \leq p \leq \tilde{p} + \sqrt{2\ln t/n}$ held with the probability of $1 - \frac{2}{T^4}$. For each time, we let $p = \tilde{p} + \sqrt{2\ln t/n}$, which exactly is the *Upper Confidence Bound*(UCB).

Appendix B

Q: For *Gradient method*, why $H_t(a)$ can work?

My Understanding For every action, $H_t(a)$ is updated by the following formula.

$$\begin{aligned} H_{t+1}(A_t) &\doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)(1 - \pi_t(A_t)), \quad for\ A_t \\ H_{t+1}(a) &\doteq H_t(a) + \alpha(R_t - \bar{R}_t)\pi_t(a)), \quad for\ a \neq A_t \end{aligned} \tag{2}$$

where $\bar{R}_t$ represents the average reward. I think it acts as a baseline. The agent choose the current action and get a reward R_t. If $R_t > \bar{R}_t$, $H_{t+1}(a)$ should grow up, or it should decline. The step size is controlled by α. Just as the saying goes 'Learning is like sailing against the current, if you don't advance you fall back'. But I can not understand why $1 - \pi_t(A_t)$ when updating the $H_{t+1}(A_t)$, so I proved it in next section.

Mathematical derivation Because I can not understand it well, it is necessary for me to prove it.

In the gradient ascent algorithm, we have

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha \frac{\partial \mathbb{E}\left[R_t\right]}{\partial H_t(a)} \tag{3}$$

We know that $\mathbb{E}\left[R_t\right] = \sum_x \pi_t(x) q_*(x)$, so

$$\begin{aligned}
\mathbb{E}\left[R_t\right] &= \sum_x \pi_t(x) q_*(x) \\
&= \frac{\partial}{\partial H_t(a)} [\sum_x \pi_t(x) q_*(x)] \\
&= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)}
\end{aligned} \tag{4}$$

The B_t is the baseline. Why there B_t is ok (2)?

$$\begin{aligned}
\sum_x B_t \frac{\partial \pi_t(x)}{\partial H_t(a)} &= B_t \sum_x \frac{\partial \pi_t(x)}{\partial H_t(a)} \\
&= B_t \frac{\partial \left[\sum_x \pi_t(x)\right]}{\partial H_t(a)} \\
&= B_t \frac{\partial [1]}{\partial H_t(a)} \\
&= 0
\end{aligned} \tag{5}$$

Then we use $w(b)$ represents H_t, b represents the every possible action. We get

$$\begin{aligned}
\frac{\partial \pi_t(x)}{\partial H_t(a)} \Leftrightarrow \frac{\partial \pi(x)}{\partial w(a)} &= \frac{\partial}{\partial w(a)}[\pi(x)] \\
&= \frac{\partial}{\partial w(a)}\left[\frac{e^{w(x)}}{\sum_{b=1}^{k} e^{w(b)}}\right] \\
&= \frac{\frac{\partial e^{w(x)}}{\partial w(a)} \sum_{b=1}^{k} e^{w(b)} - e^{w(a)} e^{w(x)}}{\left(\sum_{b=1}^{k} e^{w(b)}\right)^2} \\
&= \frac{\mathbb{I}_{a=x} e^{w(x)} \sum_{b=1}^{k} e^{w(b)} - e^{w(a)} e^{w(x)}}{\left(\sum_{b=1}^{k} e^{w(b)}\right)^2} \\
&= \mathbb{I}_{a=x}\pi(x) - \pi(x)\pi(a) \\
&= \pi(x)\left(\mathbb{I}_{a=x} - \pi(a)\right)
\end{aligned} \tag{6}$$

Bringing 6 into 4, we get

$$\frac{\partial \mathbb{E}\left[R_t\right]}{\partial H_t(a)} = \sum_x \left(q_*(x) - B_t\right) \pi_t(x) \left(\mathbb{I}_{a=x} - \pi_t(a)\right) \tag{7}$$

That is why $1 - \pi_t(A_t)$ for A_t.

References

1. F. Wei, Multi-armed bandit: Ucb (upper bound confidence). `https://zhuanlan.zhihu.com/p/32356077` Accessed January 1, 2018.

2. Z. Huiwen, Gradient gambling machine algorithm. `https://zhuanlan.zhihu.com/p/54159132` Accessed September 6, 2019.