News Articles Reccomendation using UCB Algorithm

Sokratis Siganos - 2019030097

Reinforcement Learning and Dynamic Optimization - ΠΛΗ423/ΠΛΗ723

Description

For this assignment, we are asked to implement a modified version of the ucb algorithm to create a news article reccomendation system based on people's personal preferences. Our main task is to create an algorithm which achieves sublinear regret regardless of each person's personal preferences.

Data

Each user visiting our articles is either (i) male or female, and (ii) under or over 25 years old, and each user is drawn in an IID manner.

There are a total of 5 different articles and for each one of them, there is a click propability depending on the user's category. All the propabilities based on the users' preferences are denoted on this table:

	f $\geq$ 25	m $\geq$ 25	m OR f $<$ 25
p0	0.8	0.2	0.2
p1	0.6	0.4	0.4
p2	0.5	0.5	0.8
p3	0.4	0.6	0.6
p4	0.2	0.8	0.5

The algorithm cannot use as input these propabilities. It only knows the characteristics (gender and age) of the user at the current time.

Modification Proposal

In order to create a functional algorithm based on the UCB's logic, we should take into consideration an extra parameter which is the user's characteristics. Based on intuition, applying the simple format of the UCB's algorithm won't work on this case, since each user has different tastes in articles, which means that there isn't a common optimal "arm" for every user. To solve this problem, the algorithm must find the optimal article (or articles) based on the category the user belongs in. To achieve that, we need to split the time horizon into 4 different parts, one for each user category, in which the $\mu_i(t)$ (mean reward), and the $N_i(t)$ ("pulls" of each arm) will be calculated seperately for each category. Essentially, there will be 4 different algorithm instances, finding the optimal article for each user's category.

Proof of Sublinear Regret

The main difference with the original "Optimism Under Uncertainty" algorithm, is that the parameter U is introduced, which declares the user's category. Due to its existence, the horizon of the experiment is split to $|U| = 4$ different time periods, where, $\sum_{u \in U}(T_u) = T$. Each time period T_u represents the amount of times a user belonging in a specific category has visited the website. Since the users are randomly distributed, so are the time periods. The algorithm wil implement the UCB algorithm seperately for each category of users, meaning that different variables will be calculated for each category. An instance-dependent approach will be used for proving the regret's sublinearity.

For a random category U with horizon T_u, the algorithm calculates the rewards $\hat{r}_{i,u}(t)$, the total visits of

each site $N_{i,u}(t)$, the mean reward at a specific time $\hat{\mu}_{i,u}(t)$, and the algorithm's "score" $ucb_{i,u}(t)$ where u represents the random category, $t \in T_u$ and $i \in I = \{0, 1, 2, 3, 4\}$ (set of articles). The variables are calculated as:

$$\hat{\mu}_{i,u}(t) = \frac{\sum_{n=1}^{t} \hat{r}_{i,u}(n)}{N_{i,u}(t)}, ucb_{i,u}(t) = \hat{\mu}_{i,u}(t) + \sqrt{\frac{2\ln(T_u)}{N_{i,u}(t)}}$$

Next we will prove that the algorithm for a specific category achieves sublinear regret. For simplicity, the term u will be removed from symbols, but its variable will refer to a specific user category.

Suppose that the mean regret is within a confidence interval (good event):

$$|\hat{\mu}_i(t) - \mu_i(t)| \leq \sqrt{\frac{2\ln(T_u)}{N_i(t)}} \tag{1}$$

Assuming that arm i was played at round t:

$$\mu_i(t) + \sqrt{\frac{2\ln(T_u)}{N_i(t)}} \geq \hat{\mu}_i(t) \Rightarrow \mu_i(t) + 2\sqrt{\frac{2\ln(T_u)}{N_i(t)}} \geq \hat{\mu}_i(t) + \sqrt{\frac{2\ln(T_u)}{N_i(t)}} \tag{2}$$

since $\hat{\mu}_i(t)$ belongs to the confidence interval of $\mu_i(t)$. But according to the algorithm, $ucb_i(t) \geq ucb^*(t)$ which means:

$$\hat{\mu}_i(t) + \sqrt{\frac{2\ln(T_u)}{N_i(t)}} \geq \mu^* + \sqrt{\frac{2\ln(T_u)}{N^*(t)}} \geq \mu^* \tag{3}$$

Following the leftmost and rightmost part of the inequality above:

$$\mu^* - \hat{\mu}_i(t) \leq 2\sqrt{\frac{2\ln(T_u)}{N_i(t)}} \iff \Delta_i \leq 2\sqrt{\frac{2\ln(T_u)}{N_i(t)}}$$

$$\Delta_i^2 \leq \frac{8\ln(T_u)}{N_i(t)} \iff N_i(t) \leq \frac{8\ln(T_u)}{\Delta_i^2}$$

The last inequality shows that when the Δ_i is large, the algorithm won't show this article too many times. The expected Regret for a specific user category is:

$$E[R(T_u)] = P(Good) \cdot \sum_{i=1}^{K} N_i(t) \cdot \Delta_i + P(Bad) \cdot \sum_{i=1}^{K} N_i(t) \cdot \Delta_i \tag{4}$$

Since Hoeffding's inequality applies to the random variable $N_i(t)$, the propability of a "bad" event can be described as such:

$$P(Bad) = P\left(\exists i, t : |\hat{\mu}_i(t) - \mu_i(t)| > \sqrt{\frac{2\ln(T_u)}{N_i(t)}}\right) < K \cdot t \cdot T_u^{-4} < K \cdot T_u^{-3}$$

Seperating the bad part of equation (4):

$$\sum_{i=1}^{K} N_i(t) \cdot \Delta_i \leq T_u \Rightarrow P(Bad) \cdot \sum_{i=1}^{K} N_i(t) \cdot \Delta_i \leq K T_u^{-2} \to 0$$

As the horizon grows, the "bad" part becomes insignificant so we can ignore it, which means that:

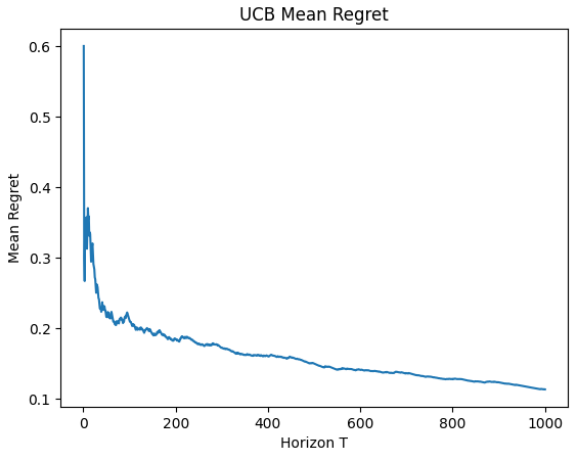
$$E[R(T_u)] = P(Good) \cdot \sum_{i=1}^{K} N_i(t) \cdot \Delta_i \leq \sum_{i=1}^{K} \frac{8\ln(T_u)}{\Delta_i} \tag{5}$$

We can summarize that this formula applies to every user's category. Since the categories are drawn in an IID manner, the total mean Regret is:

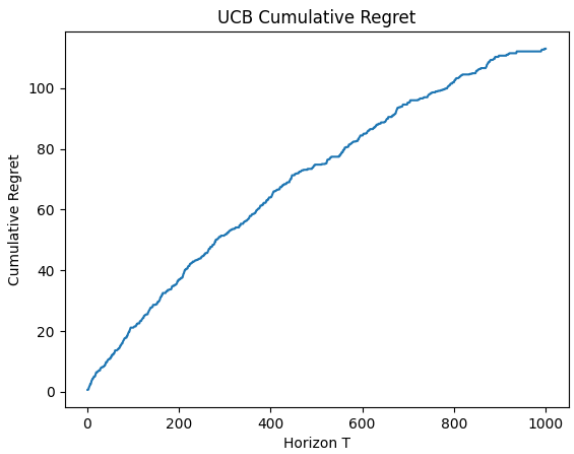
$$E[R(T)] = \sum_{\forall u} E[R(T_u)] = \sum_{\forall u} \sum_{i=1}^{K} \frac{8 \ln(T_u)}{\Delta_{i,u}} \Rightarrow$$
$$\Rightarrow E[R(T)] < \sum_{\forall i,u} \left(\frac{8K}{\Delta_{i,u}} \right) \cdot \ln(T) = O(ln(T))$$

Which proves that the algorithm's mean regret is sublinear.

Regret Plots

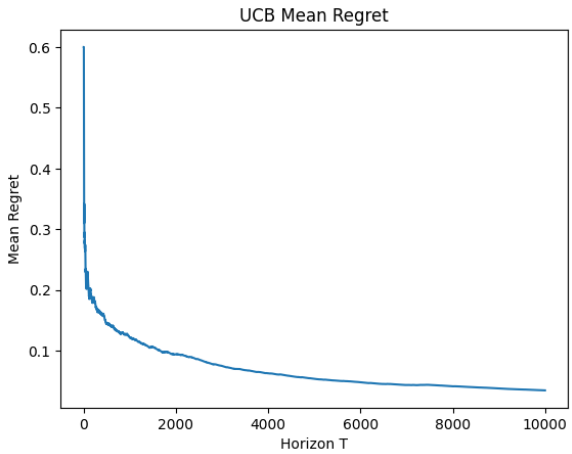


(a) Mean Regret Plot for T=1000

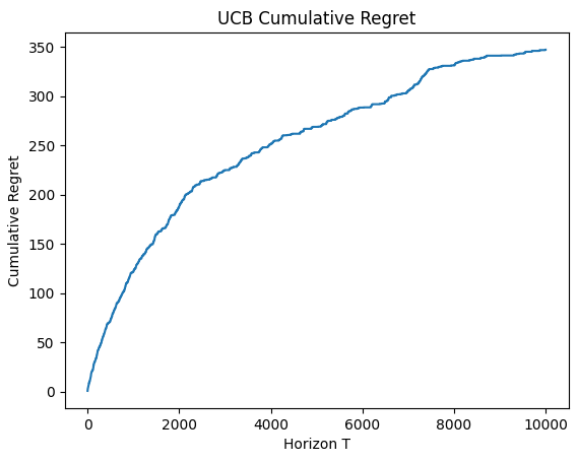


(b) Cumulative Regret Plot for T=1000

The plots above clearly show the sublinear character of this algorithm's mean regret. At first, the mean regret doesn't decrease immediately, since there are at least 20 rounds of preparation to prepare for all user types, however after the 200 round mark it clearly starts decreasing all the way down to the 0.1 mark which indicates that at this point, the algorithm strongly suggests the best article based on each user's characteristics. It is also noticeable that the slope of the cumulative regret's plot is decreasing which verifies the sublinear character of the regret.



(a) Mean Regret Plot for T=10^4



(b) Cumulative Regret Plot for T=10^4

The plots of this experiment, verify the results of the previous experiment, by showing more clearly that the mean regret is sublinear. After passing the 1000 round mark, the mean regret decreases steadily towards 0 which means that at this point most of the rounds, only the best article will be chosen. Similarly, on the cumulative regret plot, the slope keeps decreasing after the 1000 round mark.