Multi-armed Bandits and Social Networks

VVN

2019

The Problem

- We have two graphs - G_1 for the users and G_2 for the arms.
- K_1 and K_2 are the set of nodes corresponding to the graphs.
- At any time slot t (assume t is indexed for particular users), a user c (chosen uniformly from K_1) appears and is to be served by choosing an arm a $\in K_2$. For this, the recommender recieves the following information (we represent the rewards by Y):

$$Y^{c}_{t,a'} \, (\forall a' \in \mathcal{N}_2(a)) \; and \; Y^{f}_{t,a} \, (\forall f \in \mathcal{N}_1(c))$$

- The rewards are stochastic, that is , given a user and an arm , the corresponding reward follows a fixed distribution unknown to the recommender.
- The goal of the recommender is to choose arms for each user so as to maximise the respective expected reward.

Definitions and Notation

- Represent the policy of the user c by π_t^c ,which is a probabilty mass vector over all arms with $\pi_t^c(i) = Prob(Recommender\ offers\ arm\ i\ to\ user\ c\ in\ his\ t^{th}\ appearance)$
- Also, we define the vector β_t^c with :

$$\beta_t^c(i) = Prob(a_c^* = i \,|\mathcal{H}_t)$$

  where $\mathcal{H}_t$ (the "history") is the set of all rewards obtained (either directly for user c or indirectly through neighbouring users of c in the graph) before the t^{th} slot.
- We further define α_{t+1}^c as the posterior just after choosing arm a at time t :

$$\alpha_{t+1}^c(i) = Prob(a_c^* = i \,|\mathcal{H}_t, Y_{t,a'}^c\ (\forall a' \in \mathcal{N}_2(a)))$$

- We assume that we perform Thompson Sampling to decide the policy for each user. Therefore, for this special case, we have $\pi_t^c = \beta_t^c\ \forall c$.

Performance Metrics

- Define the regret as : $R_t^c = \sum_{t'=1}^{t} \mathbb{E}[Y^c_{t',a_c^*} - Y^c_{t',a} \,|\mathcal{H}_{t'}]$ with the quantity $Y^c_{t,a_c^*} - Y^c_{t,a}$ defined as instantaneous regret Δ_t^c. We wish to make this quantity converge as quickly as possible for all users.
- Define the information gain vector g_t^c as :

$$g_t^c(i) = \mathbb{E}[H(\beta_t^c) - H(\beta_{t+1}^c) \,|\mathcal{H}_{t+1}, A_t = i]$$

- For a particular policy π_t define the information ratio as :

$$\psi_t(\pi_t) = \frac{(\pi_t^T \Delta_t)^2}{(\pi_t^T g_t)}$$

- The smaller the information ratio, the more would we be sure that our policy is on the "right track", because a small value of $\psi_t \Rightarrow$ Either we are choosing actions with a small regret (exploitation) or choosing actions that would provide more expected side information (exploration).

An Example - Bernoulli rewards

- If the distribution of the rewards is assumed to be Bernoulli, and since we are doing Thompson sampling, we take the distribution of the means of the Bernoullis to be Beta with parameters$(A_{c,a}, B_{c,a})$.There would be $|K_1||K_2|$ such parameter tuples in total.
- Everytime we receive $Y_{c,a}^t = 1$ (directly as a reward or indirectly through the graph of users or arms), we increment $A_{c,a}$ by 1 and everytime we get$Y_{c,a}^t = 0$, we increment $B_{c,a}$ by 1.
- For choosing an arm for a user c , we first sample the Bernoulli means $\mu_{c,a}^t \sim Beta(A_{c,a}, B_{c,a}) \, \forall a \in K_2$ and then choose the arm $a' = argmax_a(\mu_{c,a}^t)$.

Strategy to provide bounds on Regret

- If we wish to get an upper bound on the value of regret for a Thompson Sampling based policy, we use the following theorem.

Theorem

(For this theorem ignore all side information from graphs) If $\forall t \leq T$ we know that $\psi_t \leq \psi^*$ almost surely for a particular user c, then :

$$R_T^c \leq \sqrt{\psi^* \; T \; H(\beta_1^c)}$$

Therefore, an overall strategy to provide a regret bound seems to be the following:

- Firstly, to prove an analogous theorem for the two graphs case.
- Secondly, to find a reasonably tight bound on the information ratio as a function of β_t^c and some graph properties.

A Bound on Regret

Following similar procedure used to prove the previous theorem, we get a regret bound for our case as :

Theorem

$$\mathbb{E}[R_T^c] \leq \sqrt{(\sum_{t=1}^{t=T} \mathbb{E}[\psi_t^c])H(\beta_1^c)}$$

We now need a bound on the expected information ratio $\mathbb{E}[\psi_t^c]$.

A Lower Bound on the Information Gain

We now try to find a lower bound for the information gain .

$$\because H(\beta_t^c) - H(\beta_{t+1}^c) = [H(\beta_t^c) - H(\alpha_{t+1}^c] + [H(\alpha_{t+1}^c) - H(\beta_{t+1}^c]$$

For each of the terms on the RHS, we get the following inequalities:

$$\mathbb{E}[H(\beta_t^c) - H(\alpha_{t+1}^c)|\mathcal{H}_{t+1}, A_t = i] \geq \sum_{i \in K_2} \sum_{k \in \mathcal{N}(i)} \beta_t^c(i)\, h_t^c(i)$$

$$\mathbb{E}[H(\alpha_{t+1}^c) - H(\beta_{t+1}^c)|\mathcal{H}_{t+1}, A_t = i] \geq \sum_{i \in K_2} \sum_{f \in \mathcal{N}(c)} \beta_t^f(i)\, h_t^c(i)$$

where $h_t^c(i) = I(a_c^*; Y_{c,i}|\mathcal{H}_t) = \mathbb{E}[H(\beta_t^c) - H(\beta_t^c|Y_{c,i})]$

A Lower Bound on the Information Gain (contd)

We can write the above inequalities in vector form as :

$$(\beta_t^c)^T g_t \geq (\sum_{f \in \mathcal{N}(c)} \beta_t^f)^T h_t^c + (\beta_t^c)^T (G_2 h_t^c)$$

where $G_2 = [g_2(i,j)]_{|K_2| \times |K_2|}$ is the matrix corresponding to the graph of arms.

Bounds using UCB and simulation results

- If we use the UCB algorithm instead of Thompson sampling, we obtain the following result for the upper bound on the number of times a sub optimal arm is chosen.

$$\mathbb{E}[\tau_{ij}] \leq 4\alpha\Delta_{ij}^{-2} log(t) - \sum_{i' \in \mathcal{N}_{arms}} \mathbb{E}[\tau_{i'j}] - \sum_{j' \in \mathcal{N}_{users}} \mathbb{E}[\tau_{ij'}]$$

- The lower bound for the expected number of times a sub-optimal arm is chosen is also logarithmic.
- Therefore expected regret remains of the same order whether or not we have a graph of users.
- We verify the above observation by simulation and comparing the regrets obtained in the presence and absence of the user graph.

Regret curves from the simulation

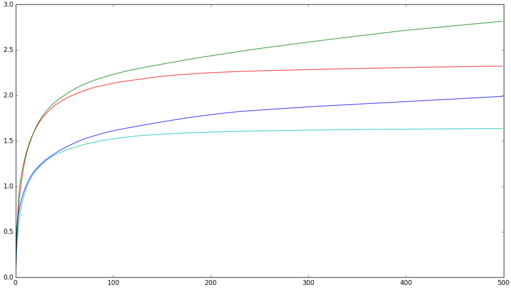

Figure 1: Cyan,Red = Thompson with and without user graph; Blue,Green = UCB with and without user graph

Regret curves from the simulation

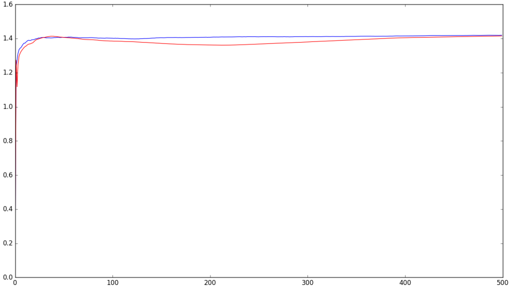

Figure 2: Red = Ratio of Thompson regret with and without user graph; Blue = Ratio of UCB regret with and without user graph

Simulation results

- It turns out that even in the case of Thompson sampling, the presence or absence of a user graph does not affect the order of the expected regret.
- Also, the constant factor by which the UCB algorithm with the user graph is better than the UCB algorithm without the graph, is the same as the factor by which Thompson is better with a graph than without.
- Therefore both UCB and Thompson do not improve anything in terms of the regret order.
- The code for the above simulation is *here*.