15.780, Fall 2021

Stochastic Models in Business Analytics

Problem Set 5 - Bandits

Due date: November 29, 2021

Instructions:

1. Submit a PDF file with your solutions to Canvas before the assigned deadline. Write your name and MIT ID on your submission.
2. All plots must have clear and easy-to-read axis labels and legends,
3. **Include relevant code in the PDF submission even if the question doesn't explicitly ask for it**. This means we can give you partial credit even if the output is wrong. if appropriate.

We highly recommend you attempt this after looking at the code from *Recitation 10*.

TeddyP needs to decide which anime show he is going to watch during finals period this year. Anime shows have hundreds of episodes, so it is a huge time commitment. He is choosing between *One Piece* and *Naruto Shipudden*. We will use multi-armed bandits (MAB) to help TeddyP decide.

The data file `anime_likes.csv` contains the likes for each episode of these shows (either a 0 or a 1). Each row corresponds to the review TeddyP would have given the show if he had watched it in that round. Of course, in reality, neither you nor TeddyP know all of this information, you would only observe the like of whichever show he chose to watch in each round (i.e. whichever arm was pulled). The full data is provided to make your coding easier, because you can just use the t'th row of the file for the reward in round t of whichever show is watched (arm is pulled).

Problem 1. (50pts)

1. Calculate the true mean reward of each show (i.e. based on all of the data in `anime_likes`). (5pts)

2. In reality of course you don't actually know the true mean rewards. Since the rewards are binary, we will use Thompson Sampling with Bernoulli rewards (like Lecture 17/Recitation 10) to estimate the shows' means. Recall that in this case we treat i's unknown mean reward μ_i as random variable following a $Beta(a_i, b_i)$, where a_i and b_i are the number of 1s and 0s we've observed for i so far (these change as we go along). Recall that, at each step, your best estimate of the true reward for arm i is $\frac{a_i}{a_i+b_i}$.

   Set the seed in R by running `set.seed(15)`. Use 300 steps of Thompson Sampling to estimate the mean reward of each anime show. Plot of your best estimate of the true means at each step vs. the step number. Include both shows on the same plot and label them clearly. (20 pts)

   Hint: You can reuse code from Recitation 10 to implement Thompson Sampling.

3. Plot the show watched (arm pulled) vs the time step for the Thompson sampling algorithm (this is like the plots you saw in lecture). (10 pts)

4. What is the estimated mean reward of each show at the end of the horizon, and how does it compare to the true mean rewards you calculated in Part 1? For each show, give a high-level explanation of why the estimate is close or far from the true mean (depending on what you see). (10pts)

5. For what fraction of the 300 rounds did you make TeddyP watch the crappier show? The crappier show is the one with the lowest reward at the end of the time horizon. (5 pts)