Introduction to Bandits

Elyn Franson

2025-06-12

Introduction

A doctor may need to have two different treatment options. Suppose I give treatment one to a patient and they recover and I give treatment two to a patient and they don't recover. How do I formalize a decision strategy to choose the better treatment for the next patients? A simple way would be to give the next patient treatment one, but are there better strategies? These sorts of problems can be related to treatments, drugs, advertisements, coin flips, etc, and are bandits problems. We looked at a problem where we had two different treatment probabilities 0.6 and 0.7 and tested if different models would make the right choice.

Greedy

The greedy method is the simplest method. Starting with the first sample we started our example with 10 trials of each drug. Next, you take the mean of both samples then you choose to select and sample the treatment with the highest mean. We sampled it once more. Then, you repeat the steps continuing to update the means each time you take a new sample.

One problem with this method is that you may end up not exploring the option with a lower sample mean. This can be inaccurate if there is a small starting sample that doesn't accurately represent the treatment mean. We looked at our problem with the probabilities of treatment 0.6 and 0.7. Then we looked at when the probabilities were further apart, 0.2 and 0.7. When the probabilities were further apart greedy was better at choosing the right treatment. After multiple trials and samplings, it chose the treatment about 100 of the time when the treatment probabilities were far apart. It got stuck when the treatment probabilities were close together and only chose the right treatment 75 of the time.

In our example above treatment 2 had no recoveries a mean of 0. If we used greedy method starting with this sample we would never choose or explore treatment 2 because the mean 0 would always be lower than the mean of treatment 1 where we had at least one success. How do we fix the problem of not exploring one of the treatments or getting stuck choosing the wrong treatment because of the beginning sample?

Epsilon Greedy

Epsilon greedy gives one solution for exploring other treatments. It chooses to sample from the treatment with the lower mean epsilon percent of the time. This helps correct the problem of getting stuck and not exploring other treatments because of a small sample size. One problem with the epsilon greedy method is that it will never fully converge to choosing the right treatment 100 percent of the time because it chooses the method with the lower mean epsilon percent of the time.

Upper Confidence Bound

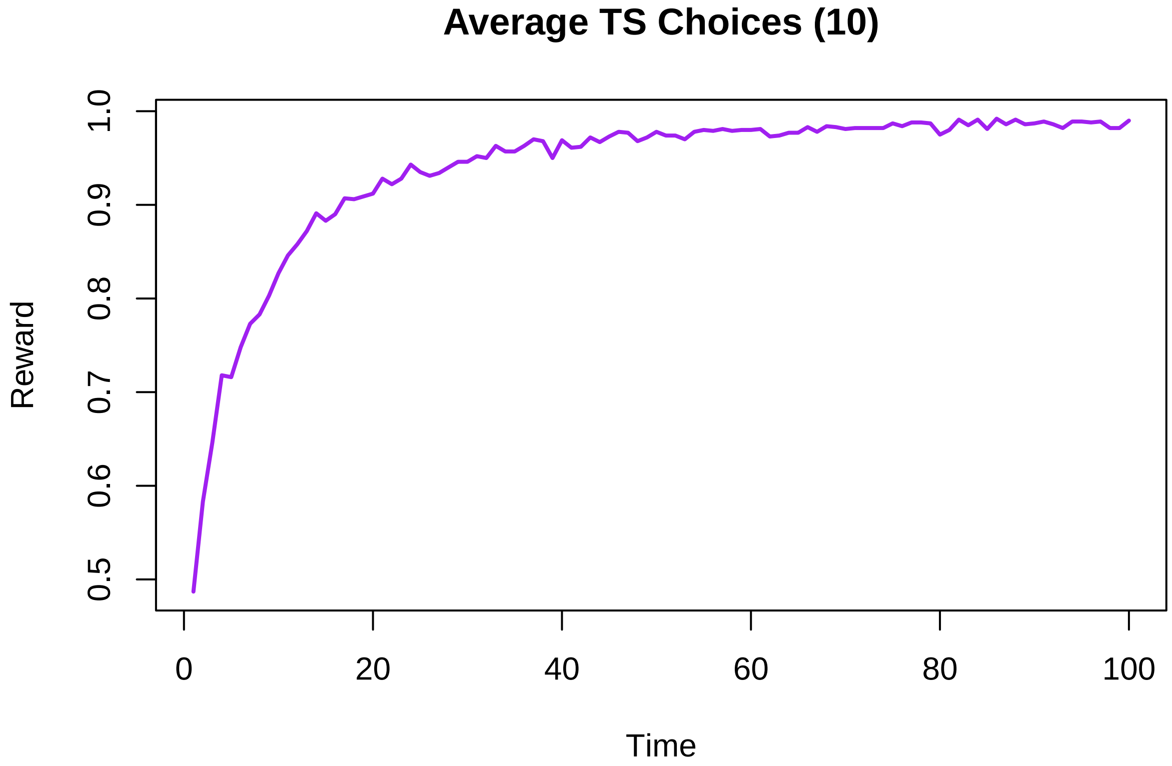
Upper confidence bound (UCB) is similar to greedy but instead of choosing the treatment with the best mean you choose the treatment with the best confidence interval. This takes into account the uncertainty of knowing a mean. With our treatment problem with probabilities 0.7 and 0.6, UCB chose the right treatment 80 of the time.

Thompson Sampling Greedy

Thompson sampling assumes a prior belief about our data. Greedy Thompson sampling starts with the prior belief and chooses to sample the one with the better mean for each distribution. We assume both treatments follow a beta distribution with α successes and β failures. The mean for the beta distribution is $\frac{\alpha}{\alpha+\beta}$ which is the success divided by total trials. This is similar to greedy in the since we are just taking the one with the best mean but we can start with a prior assumption about each distribution treatment and update our prior assumption each time a sample is taken. In our example with the same treatment probabilities, this method chose the right treatment 72 of the time. This was one of the worst methods for our example because it was like the greedy method except it started with a prior. The prior can heavily affect the outcome. If we believe a treatment has a 90 success rate it would take a lot of failures to bump it down to what may be its true success rate.

Thompson Sampling

Regular Thompson sampling improves on the greedy Thompson sampling by randomly sampling from each beta distribution. Then, it chooses the treatment with the highest mean from the random sample and updates the alpha and beta. This method converged almost to 100 of the time choosing the right method. For our problem, this was the best outcome we found.



Conclusion

In regards to solving bandits problems, the greedy method is the easiest method to implement and understand but it gets stuck or doesn't explore all treatment options. Epsilon greedy, upper confidence bound, and Thompson sampling work to solve these problems and explore all treatments. In regards to our treatment problem with the 0.6 and 0.7 probabilities of different treatments, Thompson sampling was the best option. It converged to choosing the right treatment almost 100 of the time. It randomly sampled from the beta distributions for the treatments allowing variance and choice of the right treatment.

References

Russo, Daniel J., et al. A Tutorial on Thompson Sampling. Foundations and Trends® in Machine Learning, vol. 11, no. 1, 2018, pp. 1–96. https://doi.org/10.1561/2200000070.