COMP 138 RL: 10-Arm Bandit

Tung Pham

September 26, 2024

1 Introduction

Living in the Boston Metro areas for a long time, everybody would have experience the commuting hell into the work in the area. I, as a commuting student, experience the same thing on every lecture day. Deciding which commuting means, which route to take, what bus number to avoid, how safe is the route that I'm choosing, etc. are always the deciding factor for me to decide when leaving the house. But not all the choices are always guarantee to be my best option to take due to some unexpected construction on the road, or the number of people moving into the areas unexpectedly increase for a following month causing driving tedious and having to switch to using the train, etc which extend my commuting time and causing me late to the lecture and missing out the material discussed. This issue is very much similar to the Multi-arms bandit problem.

In the traditional Multi-arm bandit problem, the robber is presented with multiple machine, each giving random money or reward that follow each machine's distribution. The goal of the robber is to determine which machine to pull in order to maximize the profit gain. Similarly, in my commuting problem, each route that I'm taking represent each arm or each machine in the bandit problem. What I'm maximizing is the amount of satisfaction on commuting based on my rating of multiple criterias.

In practice, this satisfaction can change over time based on a variety of factor and not following a single distribution due to unexpected traffic or accident. This change in amount of time saved, or expected rewards, create uncertainty and ripples in the environment space for each time step and when relating to the multi-arm bandit problem, we say that this problem is nonstationary.

2 Problem

To make the problem fit into the context of Multi-arm Bandit problem, I'll assume that I, myself, as a person that just moved to the area without any prior knowledge of the place and how much ideal time saved, denotes $Q(A)$, for each route that I'm taking. I'll record the time that I'll go to class for 10,000 times and assuming that I have 10 possible routes to take. This is a variant of multi-arm bandit where the number of arm is 10 instead of an unknown k value.

Because I've never been on any of these routes before, I don't know how much time I'm saving. Therefore, my initial $q_*(A) = 0$. I'll assume that my optimal saving time of each route I select is an incrementation that follows a random values in normal distribution that has mean of 0 and standard deviation of 0.01 which can be expressed as:

$$q_*(a, t) = q_*(a, t - 1) + X$$

where t is the current nth time I go out, and $X \sim \mathcal{N}(\mu, \sigma^2)$.

The satisfaction will be recorded as my rating based on the timeliness, the comfort, the amount of transit and the walking distance. This allow us to assume the Gaussian distribution of the rewards with the mean of 0 and standard deviation of 0.01 that is if we do the normalization of the rating correctly.

In this experiment, we'll be doing a 2000 runs, assuming moving to 2000 different town and start all over again, and then take the average amount at each time step. We'll be discussing different approach to solve this bandit problem and present visualization on how well each method were able to produce.

3 Methodology

3.1 Sample Average Action-Value Method

The first method amongs the reinforcement learning methods for the multi-arm bandit problem is the Sample Average Action-Value Method where we estimate the value of each action by taking the average of values observed for that specific action taken. This creates a general observation on how well that specific route over the course that we've lived in that region. For each route that we take, denoted $q_*(A, t)$, there will be an optimal route, denoted $Q_t(A)$ that can yield the best out come which may or may not be the route that we've taken. I'll denoted the satisfaction that or the reward to be $R_t(A)$ and the total number of times that we've picked a specific route to be $N(A)$.

Initially, since we haven't taken any action, the value of all actions should be the same and be 0 since there's no metrics recorded yet and count is 0.

$$\forall x \in A : q_0(x) = 0, N_0(x) = 0$$

Then, the estimation fucntion of each action can be written as:

$$q_*(A) = q_{t-1}(A) + \frac{1}{N_t(A)}[R_t(A) - q_{t-1}(A)]$$

As t increases, the average difference between the optimal reward and the selected reward is added to the value of the current route selection. This will curve the action value and converges to its optimality. This helps the estimation to eventually achieve optimal solution which is the true value.

The route selection or action selection is an ϵ-greedy selection with $\epsilon = 0.1$. This means that for every time step or every time I decided to go out, I'll choose an optimal selection with a 10% of the time I picked a random selection to explore with an uniformly distributed probability of selection amongst the routes.

Eventually, an average of all the runs at each step are taken by divided the accumulative rewards by the total number of runs at each time step. At the same time, the percentage that the algorithm chose to take was recorded by taking the number of times the algorithm taking the optimal selection over the total number of selections that the algorithm took at that time step...

The graph of both results will be present in the result section below.

3.2 Constant Step-size Action-Value Method

This method is similar to the Average Sample method discussed previously. However, at each time step, instead of the taking the average of the reward difference, an α step-size was provided to guide the value function towards optimality.

We can express the above using the following formula:

$$q_*(A) = q_{t-1}(A) + \alpha[R_t(A) - q_{t-1}(A)]$$

The idea of this method is that we'll increase the important of the most recent activity rather than the older ones. Since with average example method, this value is divided by the number of the times the action was selected. As the number of times we've chosen this action, the In our case, we use 0.1 as a variable for experiment.

4 Conclusion

"I always thought something was fundamentally wrong with the universe" [**?**]