INTRODUCTION TO REINFORMENT LEARNING

MULTI ARMED BANDITS

--Shubhi Agarwal

Considering an example for better understanding purpose if there is a man and he is send to an isolated place where he has to live for 300 days and the place has only 3 restaurants providing 10,8 and 5 happy meals respectively with standard deviation of 5,4 and 25. Now the man needs to know that which particular restaurant will be providing the best possible meal. For this optimal condition is when all the information for mean and standard deviation for every restaurant is already given but this will not be the case.

Here comes the **Exploration and Exploitation dilemma**. The dilemma is due to incomplete information. You need to gather enough information to make the best decisions as a whole while managing risk. When it comes to exploitation, we use the best options we know. Exploration takes certain risks in collecting information about unknown options.

Total Exploration would not lead to less regret (difference of exploration result to optimal result) as the amount of regret in this case for the above problem will be 700. Exploration of restaurants can be done by dividing the number of days with the available number of restaurants present in the city i.e. 300/3=100 then for each 100 days the man will go to one restaurant and will be exploring the meals. After exploring the man with the amount of happy meals will be 100*10+100*8+100*5 =2300 but the optimal case is 3000 so the regret is coming out to be 700 which is large enough. Hence this comes out to be a naïve approach and is bad to implement.

Total Exploitation would be the better approach as exploration will be done only for one particular time and then according to the result achieved the rest will be handled by only exploitation. For the example, if the man does the exploration for 3 days as 3 restaurants are there and obtains the result as Restaurant 1 provides him 7 happy meals, Restaurant 2, 8 happy meals and Restaurant 3 provides 5 happy meals. Then according to this the man gets to know that R2 will provide more happy meals and for rest 297 days he exploits the theory. In this particular case, regret calculated is 330(3000-2396(1*7+1*8+1*5+297*8)). Hence this has much less regret but only one time exploration may also be not accurate.

ε- GREEDY ALGORITHM: In this ε times exploration is done while 1-ε times exploitation is done. The action taken for the exploiting by averaging the past experiment outcome can be taken as:

$$Q_t(a) = \frac{\sum_{i=1}^{t} 1(a_i=a).R_i}{\sum 1_{(a_i=a)}}$$

Here the estimate for Q can be calculated at a particular time t and then max(Q_t(a)) can be calculated.

$$\begin{aligned} &\arg\max Q_t(a) \longrightarrow 1-\varepsilon \\ &\text{uniformely from } A \longrightarrow \varepsilon \end{aligned}$$

For the example if the ε value is 10% then 10% days of the total the man will explore and the other 90% he will exploit i.e. 30 days he will go for all the restaurants one by one while the other 270 days the man will exploit and go to only one restaurant based on the past experience about the meal which he has gained. While performing the calculation the regret will come out to be more less around 100 (3000 – (10*10+10*6+10*4+270*10)). This may vary but comparatively it is less than pure exploration technique.

UPPER CONFIDENCE BOUNDS: Random exploration of options given may vary as it may be present that one explores a bad action which may have given better result for the first few times but the action selected for further exploitation purpose is not suited. For this we need to favour the exploration of particular actions that has a strong potential to provide optimal value for which we need UCB algorithm.

$$U_t(a) = \sqrt{\frac{2\log t}{N_t(a)}}$$

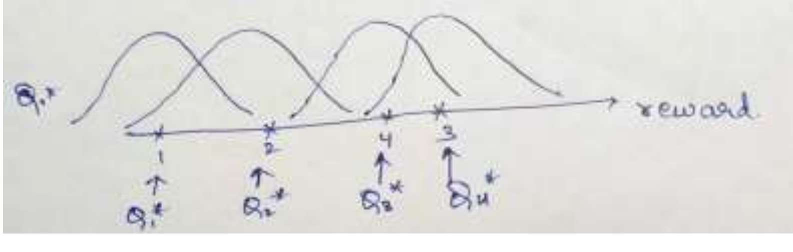
Here upper bound $U_t(a)$ is a function of $N_t(a)$, number of trials which are both inversely proportional to each other.

In UCB algorithm, we need to select the greediest action to maximize UCB:

$$a_t^{UCB1} = \arg\max_{a\in A} Q(a) + \sqrt{\frac{2\log t}{N_t(a)}}$$

In the case of the example, we cannot restrict to just having 3 restaurants and with specific deviation ratio but it may vary. For more deviation say 50% and n=3(i.e. 3 restaurants) UCB strategy wins over total exploitation while for deviation 10% vice versa is true. For the case n=10 with both deviation ratio, UCB strategy wins.

THOMPSON SAMPLING: Also called Posterior sampling. Bandits problem can be solved optimally if all its parameters are known from before. In Thompson method we will try to make assumptions about the unknown problem parameters. Assumption will be that Q* come from some distribution which can be done with Beta distribution. Now plotting all these actions on the reward line and analysing the best reward action then shaping the beta distribution along that particular point and then at each time step ‘t’ following the same rule will help in analysing the formula.



In these distribution, beta distribution of third distribution will be optimized and reach towards taking optimal decision.