RECOMMENDATION SYSTEM USING REINFORCEMENT LEARNING

MUTLI ARM BANDITS PROBLEM

-Shubhi Agarwal

B.Tech CSE 3rd year

Chandigarh University

INTRODUCTION

Reinforcement learning is a type of machine learning training methodology where machines intelligently take decisions to maximize the rewards in a particular scenario. Multi-arm bandits being one of the most common problems suggest that limited set of resources should be allocated for different alternative choices.

The problem statement focuses on building a model for a particular selling system to improvise the recommendation system so that when a customer visits then he/she is showcased only a certain amount of products according to their utility and out of those products, the salesman recommends a product which the customer will buy. We have to look into choice probabilities that a machine will refer to the customer to buy and the customer also happens to buy that product.

In handling such choices we have to take care of the exploration-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. There are many approaches to solve multi-arm bandits problem namely exploration-only , exploitation-only , ε-Greedy approach , Upper Confidence Bound(UCB) sampling and Thompson sampling.

In Total exploration, the term regret which is the difference between exploration result to the optimal result (the result which is obtained when all the conditions are previously known) is high as only exploration is just a hit and trial and doing the same for the whole time period.

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. But this approach may hinder the exploration part as if for the first time the exploration part is not the maximum one then the whole exploitation would not be able to provide the best choice. This approach but will lead to less regret than exploration-only phase. In ε- GREEDY ALGORITHM, ε times exploration is done while 1-ε times exploitation is done.

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.

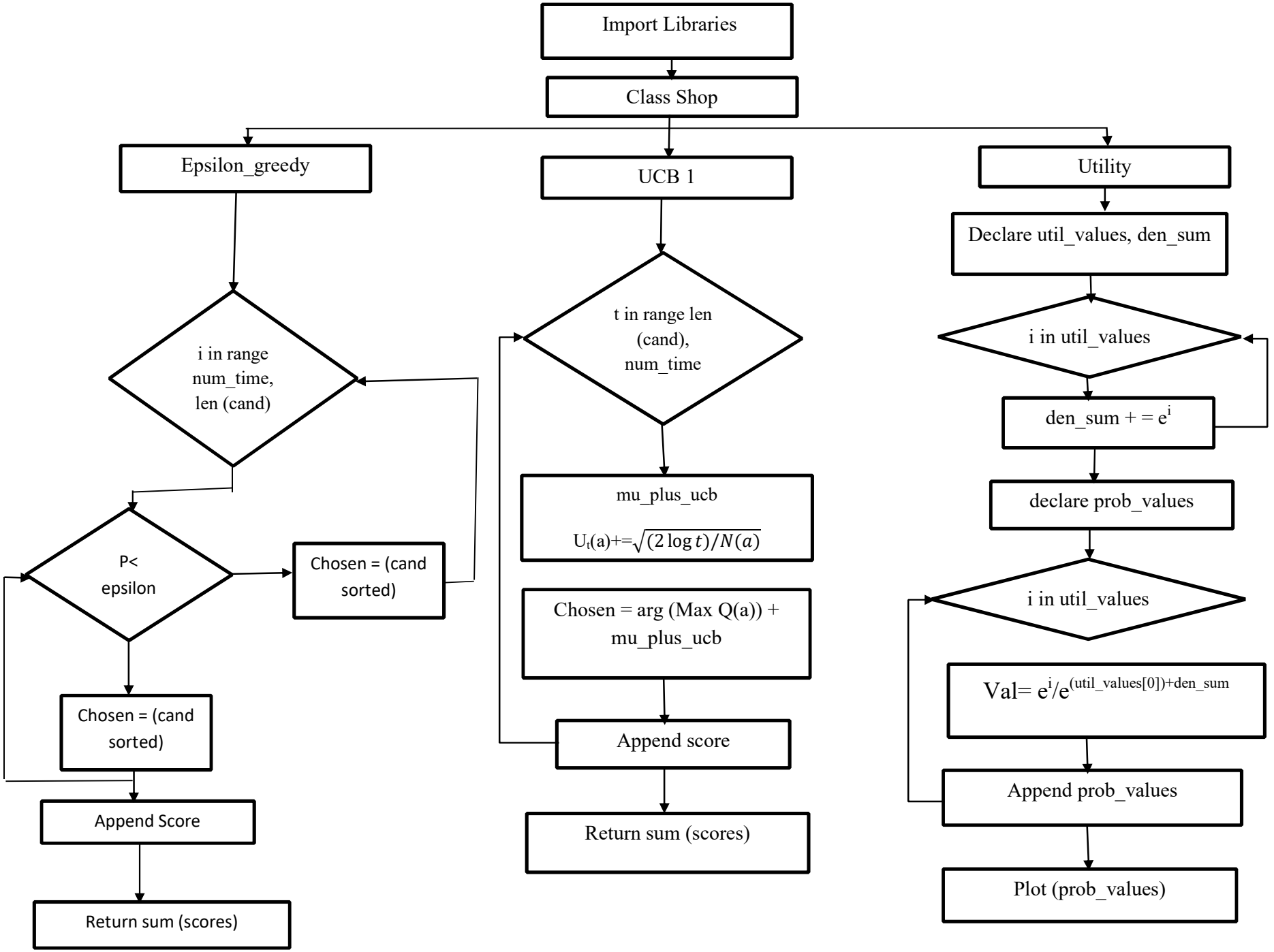
In Thompson method we will try to make assumptions about the unknown problem parameters. Assumption will be that Q* come from some distribution which can be done with Beta distribution. Now plotting all these actions on the reward line and analysing the best reward action then shaping the beta distribution along that particular point and then at each time step 't' following the same rule will help in analysing the formula.

In Multinomial Logit Modelling of supply chain management under the concept of Discrete Choice models there are three terms namely Choice set , Consumer Utility and Choice Probabilities which helps in predicting the consumer choice that he/she can make. For the discrete choice model the set of alternatives in the chice set must be mutually exclusive, exhaustive and finite where the first two requirements includes all the possible alternatives while the third requirement differentiates between discrete choice models with any other modelling like linear regression. For the second parameter i.e. Consumer Utility, the decision maker should obtain a certain level of utility from the given alternatives which is denoted as U_i where $i \in I$. Discrete choice models usually assume that the decision-maker is a utility maximizer. That is, he will choose alternative i if and only if Uni > Unj for all $j \in I$, $j \neq i$. For the purpose of error estimation in the utility values representative utilities comes into picture. The third parameter Choice Probabilities is calculated after attaining the two parameters.

METHODOLOGY

For the purpose of first analysing the problem statement we need to evaluate the regret for the different techniques for which we need to frame the functions which can be done in any computer programming language. The problem is framed as if there is a class named shop is created which is holding the value of mean values and deviation values. The flowchart will be able to display the whole functioning of the program.

S.No.	Function name	Parameters Used
1	ε-greedy	Parameters at function call: candidates, num_time, epsilon Parameters used inside function: scores[], history[],p
2	ucb1	Parameters at function call: candidates, num_time Parameters used inside function: scores[], history[],mu_plus_ucb, chosen
3	utility	util_values, den_sum, prob_values[]


Import Libraries
Class Shop
Epsilon_greedy
UCB 1
Utility
i in range num_time, len (cand)
P< epsilon
Chosen = (cand sorted)
Chosen = (cand sorted)
Append Score
Return sum (scores)
t in range len (cand), num_time
mu_plus_ucb
$U_t(a)+=\sqrt{(2\log t)/N(a)}$
Chosen = arg (Max Q(a)) + mu_plus_ucb
Append score
Return sum (scores)
Declare util_values, den_sum
i in util_values
den_sum + = e^i
declare prob_values
i in util_values
Val= $e^i/e^{(util\_values[0])+den\_sum}$
Append prob_values
Plot (prob_values)

LITERATURE REVIEW

Fundamentals of Supply Chain Theory : In this particular reading material the writer Lawrence V. Synder & Zuo-Jun Max Shen in 2019 Edition wrote about the Dicrete Choice modelling which helps in understanding purpose of how one's choices affect the chances of utilizing a particular product.

Mnl-bandit: A dynamic learning approach to assortment selection by Agrawal, S. et al. (2019) also states how to characterize the choices of assortment problems and to select the most optimal solution out of the n number of choices. The MNL-Bandit Problem: Theory and Applications by Avadhanula, V. (2019) this deals with revenue management with a significant look into customer preferences which is not at all stable and can have error no matter what the past experiences or the history of buying something depicts.

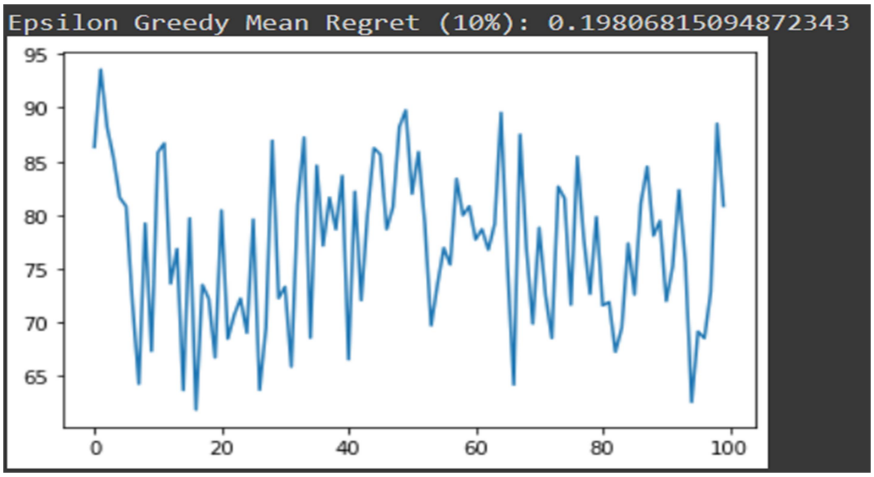
Thompson sampling for the mnl-bandit by Agrawal, S. et al. (2017) in which a subset sequential selection is done by the decision-maker and is observed in the form of the index of the MNL model with one aim of maximizing the cumulative rewards to balance the exploration-exploitation problem in the most effective way.

RESULT AND DISCUSSION

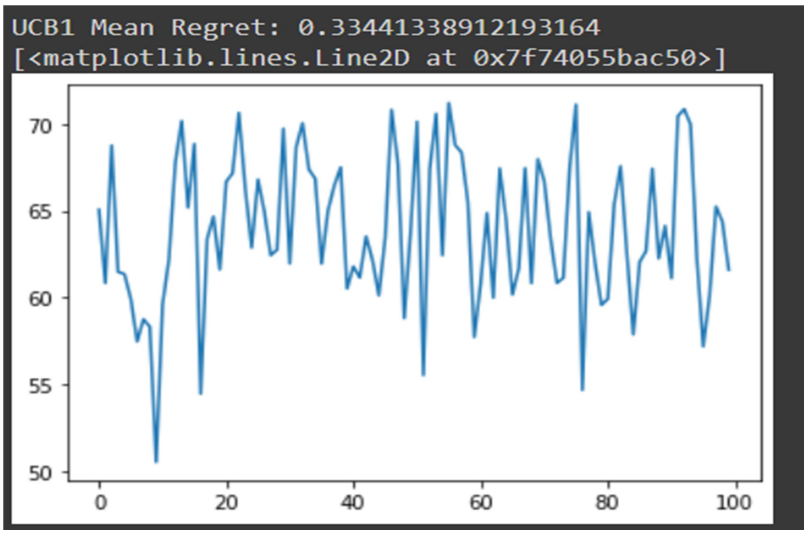
Multi-arms in the problem are the customers that arrive in a particular shop for the purpose of purchasing an item. As there are n number of products in a shop and a customer cannot buy all of them so the salesman (applying reinforcement learning) has to suggest that one product to the customer whose probability is the highest and the customer is most probable to buy that product. If a customer buys the recommended product then the recommendation system is said to provide the optimal output otherwise regret is calculated which is the amount by which the actual output may differ from the optimal output.

The problem is kept for analysing the results for 100 days in which a customer visits the shop and has some choices to choose from. A customer is thought of visiting every day and by applying different techniques of solving the exploration-exploitation dilemma. The deviation factor for this particular scenario is taken as 0.5 and number of items present in the shop is taken as 10 for the sake of simplicity.

For the epsilon value of 10% in the epsilon greedy approach the mean regret calculated for this particular scenario is 0.19806815094872343 which means around 20% of the predictions made can bear a bad result other than the optimal result and approx. 80% times the result after prediction will be optimal i.e. the customer will buy the suggested product.



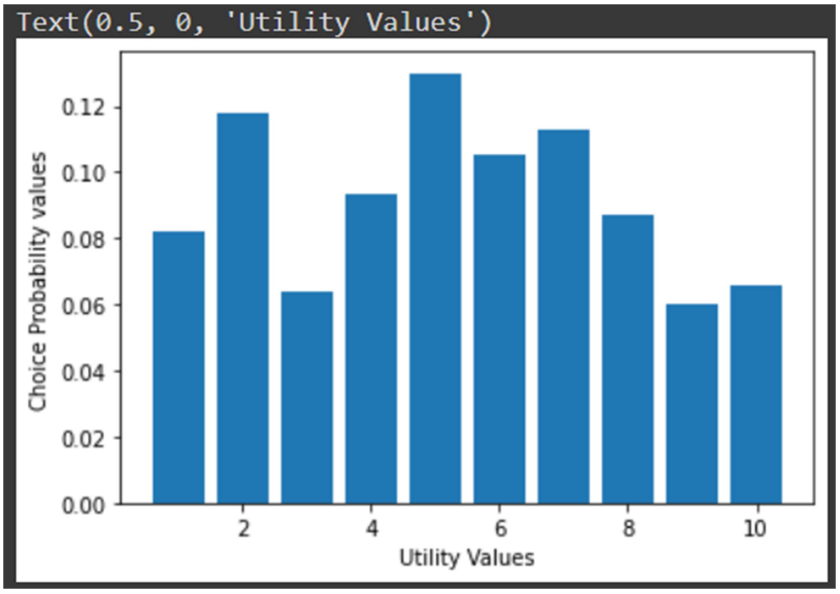
For the UCB1 approach, the mean regret is coming out to be 0.33441338912193164 which means around 33% of the predictions made can bear a bad result other than the optimal result and approx. 66% times the result after prediction will be optimal i.e. the customer will buy the suggested product.



In Supply Chain management under discrete choice models, the concept of Multinomial Logit model comes into picture where a choice set is provided to the customers where alternate choices are provided to the decision-maker and according to the utility of the customer, the probability of the choices is calculated which is the predicted probability that a customer will buy one product out of 'n' number of products. For this particular problem 10 utility values are defined and choice probability is calculated for each and every value

$$P_i(S) = \begin{cases} \frac{e^{V_i}}{e^{V_0} + \sum_j e^{V_j}}, & \text{if } i \in S \cup \{0\}, \\ 0, & \text{otherwise.} \end{cases}$$

which is then represented as the below mentioned graph:



IMPLICATIONS

The recommendation system is the most beneficial strategy in marketing area beat it maybe digital marketing or at a vendor shop. This exerts strong influence within the customers when they arrive at a particular place to buy a certain product. It also helps in generating general perceptions for a product that a customer wants to buy and when the customer will arrive for next time then this will help the vendor to keep the stock of the most buyed product more.

APPENDIX

The code of the problem statement goes like first we need to import the necessary libraries like pandas, numpy, matplotlib etc. which will help in executing the commands and visualization of data. Next a class named shop is created for storing the mean and the deviation values. There are various methodologies or approaches that can be implemented for the purpose of solving multi-arms.

In the epsilon-greedy approach for the particular epsilon value with the number of time visits and mu_dev pairs are passed as the parameters of the function. A history array is also created which helps track the past experience of the customer. Then a loop runs for number of visits – the total length of candidates inside which a local variable p is created which is compared with epsilon, if the value of p is less than epsilon then exploration is done and choices of the candidates is stored in the chosen array else the chosen is filled according to the history of the candidates choices. Now the score are appended in the scores array and sum of scores is returned from the function.

In the UCB1 approach, mean and deviation value pairs as the candidate array and number of visits is passed. Scores array and history array is created, for loop is running inside which add the values of mu_plus_ucb which is itself a function with arg(max(Q_a) + $\sqrt{(2 \log t)/N(a)}$. Then the scores are appended and the sum of the scores is returned from the function.

For the calculation of utility values we need to import math library of python firstly then the util values ray is created with the choices utilities of the customer then for the choice probabilities denominator value, denominator sum need to be calculated using for loop. Next choice probabilities array named prob_values is created now the function of choice probabilities is applied in the val variable and this will be appended in the prob_values array. Plotting of probability values is done.

For code here is the Google Colab Python Notebook :

https://colab.research.google.com/drive/1DbGalgVZSFM6Thc8XhjTXnDr2I5mJF_l?usp=sharing

https://colab.research.google.com/drive/1ocETeNVjLDvOoLj8Dc-J8XiwY4oajvEc?usp=sharing

REFERENCES

1] Lawrence V. Synder & Zuo-Jun Max Shen. Fundamentals of Supply Chain Theory Second Edition (2019) WILEY

2] Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2019). Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, *67*(5), 1453-1485.

3] Avadhanula, V. (2019). *The MNL-Bandit Problem: Theory and Applications*. Columbia University.

4] Agrawal, S., Avadhanula, V., Goyal, V., & Zeevi, A. (2017, June). Thompson sampling for the mnl-bandit. In *Conference on Learning Theory* (pp. 76-78). PMLR.