Indian Institute of Technology, Kanpur
Department of Electrical Engineering
Introduction to Reinforcement Learning (EE932)
Theory Assignment 1

Deadline: 20th April 2024 | 2023-24 Quarter 2 | Max mark: 15

- **9 objective questions followed by 3 subjective questions.**

1 Objective Questions [9 Marks]

1. We have labelled training data in
   (a) Unsupervised Learning
   (b) Supervised learning **(1)**

   Ans: (b) Supervised Learning

2. In the reinforcement learning framework, who generates the rewards for the actions taken?
   (a) Agent
   (b) Environment **(1)**

   Ans: (b) Environment

3. In the multi-arm bandits, each arm has its own underlying probability distribution from which the rewards are generated.
   (a) True
   (b) False **(1)**

   Ans: (a) True

4. Consider a 3-arm bandit ($K = 3$) problem with arms a, b, c. In Explore-Then-Commit algorithm, which arm will be played at round t=10, if each arm is explored thrice ($N = 3$) in the order $a, a, a, b, b, b, c, c, c$ and the observed rewards are $1, 2, 3,\ \ 4, -4, 3,\ \ 9, -10, 2$.

(a) Arm a

(b) Arm b

(c) Arm c

(d) A random arm is chosen **(1)**

Ans: (a) Arm a

Explanation: $\bar{\mu}(a) = 2,\ \bar{\mu}(b) = 1,\ \bar{\mu}(c) = 0.33$. Hence, arm a is the best arm based on these estimates, which will be played forever from time $t = 10$.

5. Repeat the above problem for an epsilon greedy algorithm, assuming the first 9 rounds of data are the same as above. Consider $\epsilon = 0.3$. At the round $t = 10$, what is the probability of playing arm a?

(a) 0.7

(b) 0.8

(c) 0.1

(d) 0.3 **(1)**

Ans: (b) 0.8

Explanation: $\bar{\mu}(a) = 2,\ \bar{\mu}(b) = 1,\ \bar{\mu}(c) = 0.33$. Hence, arm a is the best arm based on these estimates. A coin with the probability of heads 0.3 is tossed. If it lands in the head, exploration is done. i.e., an arm is chosen uniformly at random. Hence, due to this exploration, there is a 0.1 ($i.e., \frac{0.3}{3}$) probability of choosing arm a. Also, if the coin lands in tail, exploitiaton is done. Hence arm a will be played in that event which has a probability of 0.7. Therefor the total probability of picking arm a is 0.8.

6. The first 4 rounds of a UCB algorithm are (a, +1), (b, +2), (c, +3), (c, +2). Which arm will be picked in the 5th round? Assume $T = 1000$. Use natural logarithm (with base e) for calculations.

(a) Arm a

(b) Arm b

(c) Arm c

(d) Any arm at random. **(1)**

Ans: (b) Arm b

Explanation: The sample averages are $\bar{\mu}(a) = 1,\ \bar{\mu}(b) = 2,\ \bar{\mu}(c) = 2.5$.

Number of plays till previous (4th) round are $n_{t-1}(a) = 1, n_{t-1}(b) = 1, n_{t-1}(c) = 2$.

Consideringr the UCB scores, arm b, which has the highest UCB score, will be played in the 5th round.

- $\bar{\mu}(a) + \sqrt{\frac{2 \log T}{n_4(a)}} = 1 + 3.7169$
- $\bar{\mu}(b) + \sqrt{\frac{2 \log T}{n_4(b)}} = 2 + 3.7169$
- $\bar{\mu}(c) + \sqrt{\frac{2 \log T}{n_4(c)}} = 2.5 + 2.6282$

7. Let there be two arms a, b. Assume total round $T = 5$. Assume that the true means (which are unknown to the agent) of the arms be $\mu(a) = 4$, $\mu(b) = 3$. If I follow an algorithm that picks arms uniformly at random in all the rounds, what is the expected regret performance of that algorithm? $T\mu^* - \sum_{t=1}^{T} \mathbb{E}[R_t]$

   (a) 3.5

   (b) 2.5

   (c) 15

   (d) 5 **(1)**

Ans: (b) 2.5

Explanation:

- Based on the true means, the best expected reward is $\mu^* = \max\{\mu(a), \mu(b)\} = 4$.
- Hence, if the best arm is known, the best expected cumulative reward would be $T\mu^* = 5 \times 4 = 20$.
- The randomized algorithm is picking both the arms with equal probability. Hence, the expected reward per round (at any round t) would be $\mathbb{E}[R_t] = 4 \times \frac{1}{2} + 3 \times \frac{1}{2} = 3.5$. Hence the total expected reward of this algorithm is $5 \times 3.5 = 17.5$, which results in a regret of 2.5.

.

8. In the epsilon-greedy algorithm, should we consider gradually increasing/decreasing the value of ϵ as a function of our current round t?

   (a) Gradually increase the value of ϵ

   (b) Gradually decrease the value of ϵ **(1)**

   Ans: (b) Gradually decrease

   Explanation: As time progresses, we get to see more and more samples of rewards, and therefore, our confidence in the estimate estimated means increases. Hence, the amount of exploration can be decreased as time t increases.

9. When we use clustering technique to solve the conextual bandits, we assume all the user in a cluster to have same expected rewards.

   (a) True

   (b) False **(1)**

   Ans: True

2 Subjective Questions [6 Marks]

1. Consider a contextual bandits scenario in which the true mean $\mu_a(x) = {\theta_a}^T x$ of an arm a is a linear function of the context vector x. Here θ_a and x are $n \times 1$ vectors if n is the number of features in the context vector. Assume that we have two arms a_1 and a_2, and the samples *(context, action, reward)* observed by the agent in the first 6 rounds are as follows:

$$(\begin{bmatrix}1\\3\end{bmatrix}, a_1, r = 17),$$
$$(\begin{bmatrix}7\\13\end{bmatrix}, a_2, r = 2),$$
$$(\begin{bmatrix}5\\7\end{bmatrix}, a_1, r = 2),$$
$$(\begin{bmatrix}5\\3\end{bmatrix}, a_2, r = 1),$$
$$(\begin{bmatrix}11\\13\end{bmatrix}, a_1, r = 23),$$
$$(\begin{bmatrix}5\\7\end{bmatrix}, a_2, r = 9)$$

If the context seen in the 7^{th} round is $\begin{bmatrix}2\\1\end{bmatrix}$, what arm is played by the agent in that round if it uses an ETC policy? Write down your answer and upload it. (2)

Ans: Arm a_2

Solution: Let us form the feature matrix D_a and reward vector b_a for both the arms based on the data and rewards in the exploration phase

$$D_{a_1} = \begin{bmatrix}1 & 3\\5 & 7\\11 & 13\end{bmatrix}, b_{a_1} = \begin{bmatrix}17\\2\\23\end{bmatrix}$$

$$D_{a_2} = \begin{bmatrix}7 & 13\\5 & 3\\5 & 7\end{bmatrix}, b_{a_2} = \begin{bmatrix}2\\1\\9\end{bmatrix}$$

Given the data, let us estimate θ_a for both the arms based on the following equation

$$\hat{\theta}_a = (D_a^\top D_a + I)^{-1} D_a^\top b_a$$

here, $I = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$

Solving for $\hat{\theta}_{a_1}$ and $\hat{\theta}_{a_2}$ we get:

$$\hat{\theta}_{a_1} = \begin{bmatrix}-2.08\\3.25\end{bmatrix}, \hat{\theta}_{a_2} = \begin{bmatrix}0.56\\0.06\end{bmatrix}$$

Now, in the 7^{th} round, the feature vector of the new user is $x = \begin{bmatrix}2\\1\end{bmatrix}$. Estimating mean return for each arm $\hat{\mu}_a = x^\top \theta_a$ based on the given context feature x

$$\hat{\mu}_{a_1} = 2 * (-2.08) + 1 * (3.25) = -0.91$$
$$\hat{\mu}_{a_2} = 2 * (0.56) + 1 * (0.06) = 1.17$$

Since, $\hat{\mu}_{a_2} > \hat{\mu}_{a_1}$ hence arm a_2 will be played in the 7^{th} round.

2. If LinUCB algorithm is used, what are the UCB scores of arm a_1 and arm a_2 for the above problem w.r.t to the context seen in the 7th round? Write down your answer and upload it. (2)

Ans: Arm a_2

Solution: Calculating UCB_a score for each arm based on the following

$$UCB_a = \underbrace{x^\top \hat{\theta}_a}_{\text{Exploitation Score}} + \underbrace{\sqrt{x^\top (D_a \top D_a + I)^{-1} x}}_{\text{Exploration Score}}$$

For arm a_1:

$$UCB_{a_1} = -0.91 + 0.58 = -0.33$$

For arm a_2:

$$UCB_{a_2} = 1.17 + 0.39 = 1.56$$

Since, $UCB_{a_2} > UCB_{a_1}$ hence arm a_2 will be played in the 7^{th} round.

3. In the contextual bandits, we have used a feature vector to represent users since there are too many users to handle individually. Now consider a case where the number of arms is also too large. Do you have any suggestions for handling this situation? Please type your answer and upload it. (2)

   Ans: The following techniques can be used to handle a large action space:

   1. **Clustering of arms**: Any arms may be chosen from the cluster
   2. **Function approximation**: For handling large number of users we used a feature vector x for each user. Similarly, for handling large number of arms we used a feature say p for each arm. The mean of the arm a_i with feature vector p_i for a user u_j with feature x_j can be approximated using a function of these feature vectors i.e. $\mu_{a_i} = f(x_j, p_i)$. One way to learn this function would be to use the ETC algorithm using the data generated during the exploration phase. This learned function can then be used to predict the expected mean for making decisions during the subsequent exploitation phase.