EE932 Assignment-1 Solution

eMasters in Communication Systems, IITK EE932: Introduction to Reinforcement Learning Instructor: Prof. Subrahmanya Swamy Peruru Student Name: Venkateswar Reddy Melachervu

Roll No: 23156022

Question 10: Consider a contextual bandits scenario in which the true mean $\mu(\bar{x}) = \theta_a^T \bar{x}$ of an arm a is a linear function of the context vector \bar{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 samples (context, action, rewards) observed by the agent in the first 6 rounds as follws:

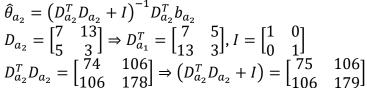
$$\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}, a_1, r = 17 \right), \left(\begin{bmatrix} 7 \\ 13 \end{bmatrix}, a_2, r = 2 \right), \left(\begin{bmatrix} 5 \\ 7 \end{bmatrix}, a_1, r = 2 \right), \left(\begin{bmatrix} 5 \\ 3 \end{bmatrix}, a_2, r = 1 \right), \left(\begin{bmatrix} 11 \\ 13 \end{bmatrix},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

If the context seen in 7^{th} round is $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, what arm is played by the agent in that round if it uses ETC policy? Upload an attachment showing your solution.

Solution:

For ETC:

- Explore each arm 2 times
- The context given is for round 7 and we need to find which is the arm to be played for this round sing ETC.
- In ETC, the arm to be played in round t=7>NK (N is exploration rounds and K is number of arms) is $a_7=\arg\max_{a=(a_1,a_2)}\hat{\theta}_a^T\bar{x}^7$
- Here we have K=2 and given two features per context, each arm is a 2-d vector and to estimate the two dimensional paramters of $\hat{\theta}_a$, we need at least two samples $\Rightarrow N=2\times 2=4$
- ullet Let's consider the first two samples of each arm for estimating $\widehat{ heta}_a$ for both the arms
- Estimate for using Ridge regression for each arm $\hat{\theta}_a = (D_a^T D_a + I)^{-1} D_a^T b_a$
 - \circ Where D_a is 2 imes 1 is a context vector with each row representing feature vectors of the each arm
 - \circ b_a is 2×1 reward vector with rewards obtained during 2 exploration rounds of the respective arm





Let's compute $\hat{\theta}_{a_1}^T \bar{x}^7$, $\hat{\theta}_{a_2}^T \bar{x}^7$

$$\mu(a_1) = \widehat{\theta}_{a_1}^T \overline{x}^7 = \begin{bmatrix} -5.883 & 4.880 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = -6.8792$$

$$\mu(a_2) = \widehat{\theta}_{a_2}^T \overline{x}^7 = \begin{bmatrix} 0.166 & 0.074 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 0.3723$$

 $\mu(a_2) > \mu(a_1)$

 \therefore Arm a_2 will be played in 7th round

----- End of the Document -----