EEE932 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 7th 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 $\Rightarrow N = 2 \times 2 = 4$
- Let's consider the first two samples of each arm for estimating $\hat{\theta}_a$ for both the arms
- Estimate for using *Ridge regression* for each arm - $\hat{\theta}_a = (D_a^T D_a + I)^{-1} D_a^T b_a$
    - Where D_a is 2×1 is a context vector with each row representing feature vectors of the each arm
    - b_a is 2×1 reward vector with rewards obtained during 2 exploration rounds of the respective arm

$$\hat{\theta}_{a_1} = \left(D_{a_1}^T D_{a_1} + I\right)^{-1} D_{a_1}^T b_{a_1}$$

$$D_{a_1} = \begin{bmatrix}1 & 3\\5 & 7\end{bmatrix} \Rightarrow D_{a_1}^T = \begin{bmatrix}1 & 5\\3 & 7\end{bmatrix}, I = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$

$$D_{a_1}^T D_{a_1} = \begin{bmatrix}26 & 38\\38 & 58\end{bmatrix} \Rightarrow \left(D_{a_1}^T D_{a_1} + I\right) = \begin{bmatrix}27 & 38\\38 & 59\end{bmatrix}$$

$$\left(D_{a_1}^T D_{a_1} + I\right)^{-1} = \frac{1}{149}\begin{bmatrix}59 & -38\\-38 & 27\end{bmatrix}$$

$$D_{a_1}^T b_{a_1} = \begin{bmatrix}1 & 5\\3 & 7\end{bmatrix}\begin{bmatrix}17\\2\end{bmatrix} = \begin{bmatrix}27\\65\end{bmatrix}$$

$$\boldsymbol{\hat{\theta}_{a_1}} = \frac{1}{149}\begin{bmatrix}59 & -38\\-38 & 27\end{bmatrix}\begin{bmatrix}27\\65\end{bmatrix} = \begin{bmatrix}0.396 & -0.255\\-0.255 & 0.181\end{bmatrix}\begin{bmatrix}27\\65\end{bmatrix} = \begin{bmatrix}\mathbf{-5.883}\\\mathbf{4.880}\end{bmatrix}$$

$$\hat{\theta}_{a_2} = \left(D_{a_2}^T D_{a_2} + I\right)^{-1} D_{a_2}^T b_{a_2}$$

$$D_{a_2} = \begin{bmatrix}7 & 13\\5 & 3\end{bmatrix} \Rightarrow D_{a_1}^T = \begin{bmatrix}7 & 5\\13 & 3\end{bmatrix}, I = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$

$$D_{a_2}^T D_{a_2} = \begin{bmatrix}74 & 106\\106 & 178\end{bmatrix} \Rightarrow \left(D_{a_2}^T D_{a_2} + I\right) = \begin{bmatrix}75 & 106\\106 & 179\end{bmatrix}$$

$$\left(D_{a_2}^T D_{a_2} + I\right)^{-1} = \frac{1}{2189}\begin{bmatrix} 179 & -106 \\ -106 & 75 \end{bmatrix}$$

$$D_{a_2}^T b_{a_2} = \begin{bmatrix} 7 & 5 \\ 13 & 3 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 19 \\ 29 \end{bmatrix}$$

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{a_2}} = \frac{1}{2189}\begin{bmatrix} 179 & -106 \\ -106 & 75 \end{bmatrix}\begin{bmatrix} 19 \\ 29 \end{bmatrix} = \begin{bmatrix} 0.082 & -0.048 \\ -0.048 & 0.034 \end{bmatrix}\begin{bmatrix} 19 \\ 29 \end{bmatrix} = \begin{bmatrix} \mathbf{0.166} \\ \mathbf{0.074} \end{bmatrix}$$

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

$$\boldsymbol{\mu(a_1) = \widehat{\theta}_{a_1}^T \bar{x}^7 = [-5.883 \quad 4.880]\begin{bmatrix} 2 \\ 1 \end{bmatrix} = -6.886}$$

$$\boldsymbol{\mu(a_2) = \widehat{\theta}_{a_2}^T \bar{x}^7 = [0.166 \quad 0.074]\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 0.406}$$

$$\boldsymbol{\mu(a_2) > \mu(a_1)}$$

∴ **Arm $\boldsymbol{a_2}$ will be played in 7th round**

-- End of the Document --

"… dare to dream; care to win …" ©

C-501, Salarpuria Serenity, 5th Main, Sector 7, HSR Layout, Bengaluru 560102 KA India
Mobile: +91 97012 22130, Email: vmelachervu@gmail.com, vmela23@iitk.ac.in, Website: www.linkedin.com/in/vmelachervu