EE675A : Introduction to Reinforcement Learning

Midsem practice solutions

04/05/2023

Lecturer: Subramanya swamy peruru *Scribe: A.V.Jayanth reddy*

1 Contextual bandits

Q1: Given an over-determined system of linear equations $Ay = b$, where there are more equations than unknowns, assuming $(A^TA)^{-1}$ exists, we generally use the concept of pseudo-inverse to compute $\hat{y} = (A^TA)^{-1}A^Tb$ as the estimate of y. Find $\hat{y}$ if

$$A = \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix}, b = \begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}$$

A1: To compute the pseudo-inverse of a matrix, we can use the formula: $\hat{y} = (A^TA)^{-1}A^Tb$

$$(A^TA) = \begin{bmatrix} 147 & 181 \\ 181 & 227 \end{bmatrix}$$

$$(A^TA)^{-1} = \begin{bmatrix} 0.3734 & -0.2977 \\ -0.2977 & 0.2418 \end{bmatrix}$$

$$(A^TA)^{-1}A^T = \begin{bmatrix} -0.5197 & -0.2171 & 0.2368 \\ 0.4276 & 0.2039 & -0.1316 \end{bmatrix}$$

from this, we can estimate the value of $\hat{y}$ to be $\begin{bmatrix} -7.513 \\ 8.11 \end{bmatrix}$

Q2: 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 7th round is $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, what arm is played by the agent in that round if it uses a greedy policy?

A2: Given the samples (context, action, reward) observed by the agent in the first 6 rounds, we can estimate the parameters a using linear regression. For each arm, we can collect the samples where that arm was selected and use them to estimate the corresponding parameters. Specifically, we can use the least squares method to estimate θ_a as follows:

$\theta_a = (Xa^T Xa)^{-1} Xa^T ra$

for a_1 , $Xa_1 = \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix}$, $ra_1 = \begin{bmatrix} 17 \\ 2 \\ 23 \end{bmatrix}$

for a_2 , $Xa_2 = \begin{bmatrix} 7 & 13 \\ 5 & 3 \\ 5 & 7 \end{bmatrix}$, $ra_2 = \begin{bmatrix} 2 \\ 1 \\ 9 \end{bmatrix}$

from this we can obtain the sample estimated to be $\theta_1 = \begin{bmatrix} -3.8 \\ 4.6 \end{bmatrix} and \theta_2 = \begin{bmatrix} 0.6 \\ 0.0 \end{bmatrix}$

To determine the arm to select in the 7th round using a greedy policy, we compute the expected reward for each arm given the observed context: $\mu_1 = \theta_1^T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = -2.8$, $\mu_2 = \theta_2^T \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 1.2$

Since the expected reward for arm 2 is higher than that of arm 1, the agent should select arm 2 in the 7th round if it uses a greedy policy.

2 Policy Gradient Algorithm

Q3: Consider a parametric representation of policy π given by

$\pi(a; \{\theta_b\}_{b \in \mathcal{A}}) = \frac{\exp(\theta_a)}{\sum_{b \in \mathcal{A}} \exp(\theta_b)}$, for $a \in \mathcal{A}$.

Here $\mathcal{A}$ is the set of actions, and $\{\theta_b\}_{b \in \mathcal{A}}$ is the parameter vector that characterizes the policy π. Assuming there are only two actions a_1 and a_2, derive the policy gradient update for this case.

A3: In the policy gradient algorithm, the policy gradient update is given by

$\theta_i^{(t+1)} = \theta_i^{(t)} + \alpha(E[R_t \frac{dlog(\pi(x))}{d\theta_i}])$

$\pi(a_i; \theta) = \frac{exp(\theta_i)}{\sum_{j \in \mathcal{A}} \exp(\theta_j)}$,

Here, $\frac{dlog(\pi(a_i;\theta))}{d\theta_k} = \frac{d(log(exp(\theta_i) - \sum_{j \in \mathcal{A}} \exp(\theta_j)))}{d\theta_k}$

$= 1_{i=k} - \frac{exp\theta_k}{\sum_{j \in \mathcal{A}} \exp(\theta_j)}$

we know that, $\pi(a_1) = \frac{exp(\theta_1)}{exp(\theta_1) + exp(\theta_2)}$ and $\pi(a_2) = \frac{exp(\theta_2)}{exp(\theta_1) + exp(\theta_2)}$

If we picked arm 1, then the update will be -

$\theta_1^{(t+1)} = \theta_1^{(t)} + \alpha(R_t)(1 - \pi(a_1))$; $\theta_2^{(t+1)} = \theta_2^{(t)} + \alpha(R_t)(0 - \pi(a_2))$

If we picked arm 2, then the update will be -

$\theta_2^{(t+1)} = \theta_2^{(t)} + \alpha(R_t)(1 - \pi(a_2))$; $\theta_1^{(t+1)} = \theta_1^{(t)} + \alpha(R_t)(0 - \pi(a_1))$