Stanford University

AA228/CS238: Decision Making Under Uncertainty

Winter 2023

Sydney Katz • 320-105 (Braun Hall) • email: *smkatz@stanford.edu*

PROBLEM SESSION 7: BANDITS AND BELIEFS

February 22, 2023 4:30pm PT

Question 1. Multi-Armed Bandits

A student in the Stanford Intelligent Systems Lab has been hard at work trying to write her PhD thesis. To ensure that it is not all left to the last minute, she has adopted the strategy of working on it for at least one hour per day. However, the time of day for her thesis-writing hour varies each day. She has two options:

1. She can write when she first gets to lab in the morning.
2. She can write when she gets home at night.

Each option results in one of the following outcomes:

1. She gets some outlining or 1-2 paragraphs done (reward of 0).
2. She has a burst of inspiration and gets 1-2 pages done (reward of +1).

Working in the morning has some unknown probability of each outcome, while working at night has a different unknown probability of each outcome. She wants to select the time to write each day to maximize her productivity. To procrastinate writing her thesis, she decides to formulate this as a multi-armed bandit problem.

a) What are the “arms” in this bandit problem? What represents a success in this case?

> **Solution:**
> The arms are the time of day for writing. There is one arm for writing in the morning, and one arm for writing at night. A success is having a burst of inspiration and getting 1-2 pages done to receive a reward of +1.

b) We want to take a Bayesian approach to modeling our belief over the probability of success for each arm. Let θ_1 represent the probability of success for writing in the morning and θ_2 represent the probability of success for writing at night. How should we represent our belief over these probabilities?

> **Solution:**
> Since we have two discrete outcomes, we should use a Beta distribution: $\theta_1 \sim \text{Beta}(\alpha_1, \beta_1)$ and $\theta_2 \sim \text{Beta}(\alpha_2, \beta_2)$.

c) If we assume a uniform prior over these outcomes, what should our prior be? What if we believe based on prior knowledge that it is more likely to get 1-2 paragraphs done than to get 1-2 pages. What would a reasonable prior be in this case?

> **Solution:**
> If we assume a uniform prior, we should use $\text{Beta}(1, 1)$. If we instead pick the prior based on the knowledge that it is more likely to get 1-2 paragraphs done than 1-2 pages, we should use a prior that starts with a higher pseudocount for failures (e.g. $\text{Beta}(1, 5)$).

d) Assume we start with a Beta$(1, 5)$ prior. We observe data for 4 mornings and 4 nights. One of the nights, the student gets 1-2 pages done. She gets 1-2 paragraphs done in all other writing sessions. What is our posterior belief over the probability of success when writing in the morning? What about when writing at night?

Solution:
In general, if we start with a prior of Beta(α, β) and observe n successes out of m trials, our posterior should be updated to Beta$(\alpha + n, \beta + m - n)$. We observed 0 successes in 4 trials of morning writing, so our posterior for morning is Beta$(1+0, 5+4-0) = \boxed{\text{Beta}(1, 9)}$. We observed 1 success in 4 trials of night writing, so our posterior for night is Beta$(1+1, 5+4-1) = \boxed{\text{Beta}(2, 8)}$.

e) Assume we are following an ϵ-greedy policy with $\epsilon = 0.2$ to decide when to write tomorrow. Our belief over the probabilities of success is based on the posteriors calculated in part (d). What is the probability that we choose to write in the morning?

Solution:
First, we should calculate the posterior probability of success ρ_a for each action. The posterior probability of success for a distribution Beta(α, β) is $\rho = \frac{\alpha}{\alpha+\beta}$.

$$\rho_1 = \frac{1}{1+9} = \frac{1}{10}$$

$$\rho_2 = \frac{2}{2+8} = \frac{2}{10}$$

The greedy action in this case is therefore to write at night. With probability $1 - \epsilon = 0.8$, we will take the greedy action. With probability of $\epsilon = 0.2$, we will select an action at random. Given that are selecting an action at random, we will have a 0.5 probability of writing at night and a 0.5 probability of writing in the morning. Thus, we will select the morning action with probability $(0.2)(0.5) = \boxed{0.1}$.

f) Consider the same scenario as in part (e) but assume instead that we want to use softmax exploration with precision parameter $\lambda = 1$. What is the probability we choose to write in the morning?

Solution:
We will select our actions with probability $\frac{\exp(\lambda\rho_a)}{\sum_a \exp(\lambda\rho_a)} = \frac{\exp(\rho_a)}{\sum_a \exp(\rho_a)}$. Thus, we will select the morning action with probability

$$\frac{\exp(0.1)}{\exp(0.1) + \exp(0.2)} = \boxed{0.475}.$$

g) Consider the same scenario as in part (e) but assume instead that we want to use the UCB1 strategy for exploration with exploration parameter $c = 1$. What is the probability we choose to write in the morning?

Solution:
When we use UCB1, we select the action that maximizes $\rho_a + c\sqrt{\frac{\log N}{N(a)}}$. Using a Bayesian approach,

N comes from the pseudocounts. Thus, $N(1) = 1 + 9 = 10$, $N(2) = 2 + 8 = 10$, and $N = N(1) + N(2) = 20$.

$$\rho_1 + c\sqrt{\frac{\log N}{N(1)}} = 0.1 + (1)\sqrt{\frac{\log 20}{10}} = 0.647$$

$$\rho_2 + c\sqrt{\frac{\log N}{N(2)}} = 0.2 + (1)\sqrt{\frac{\log 20}{10}} = 0.747$$

The UCB1 algorithm will therefore tell use to write at night, so the probability we write in the morning is $\boxed{0.0}$.

h) Instead of using the previous ad hoc exploration strategies, we decide to solve for the optimal exploration strategy noting that there is a finite number of days until the thesis is due. Let w_i represent the number of successes when taking action i ($i = 1$ for morning and $i = 2$ for night) and similarly let ℓ_i represent the number of losses. Write down an expression for computing the optimal state-action utility function for the morning action (i.e. $Q^*(w_1, \ell_1, w_2, \ell_2, \text{morning})$). Assume we use the prior from part (d) (Beta$(1, 5)$ for probability of success both in the morning and at night).

Solution:

$$\begin{aligned} Q^*(w_1, \ell_1, w_2, \ell_2, \text{morning}) =& \frac{w_1 + 1}{w_1 + \ell_1 + 6}(1 + U^*(w_1 + 1, \ell_1, w_2, \ell_2)) + \\ & \left(1 - \frac{w_1 + 1}{w_1 + \ell_1 + 6}\right) U^*(w_1, \ell_1 + 1, w_2, \ell_2) \end{aligned}$$

i) Describe in words how you would calculate the optimal state-action value function if we know that there are 30 days until the thesis is due.

Solution:
We would want to start with all states such that $\sum_a (w_a + \ell_a) = 30$ and let U^* for these states equal 0. We would then work backwards starting with states where $\sum_a (w_a + \ell_a) = 29$ iteratively applying the equation from part (h) for the morning action and a similar version for the night action until we reach the initial state.

j) After testing the strategy for a few days, the student realizes there is one more possible outcome of the one hour writing session: she gets distracted by something else and does not make any progress on her thesis for a reward of -2. For each action, we now have three probabilities to estimate. Let $\theta_1 = [\theta_{11}, \theta_{12}, \theta_{13}]$ be the probabilities for each outcome (no progress, 1-2 paragraphs, 1-2 pages respectively) for taking the morning action and let $\theta_2 = [\theta_{21}, \theta_{22}, \theta_{23}]$ be defined similarly. How should we represent our belief over θ_1 and θ_2?

Solution:
There are now three discrete outcomes each. We should represent our belief using a Dirichlet distribution (e.g. $\theta_1 \sim \text{Dir}(\alpha_{11}, \alpha_{12}, \alpha_{13})$).

k) Suppose we start with a uniform prior. Let w_{ij} represent the number of times action i resulted in outcome j ($j = 1$ for no progress, $j = 2$ for 1-2 paragraphs, and $j = 3$ for 1-2 pages). Write down the new expression for computing the optimal state-action utility function for the morning action (i.e. $Q^*(w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, \text{morning})$).

Solution:

$$\begin{aligned}Q^*(w_{11}, w_{12}, w_{13}, w_{21}, w_{22}, w_{23}, \text{morning}) =& \frac{w_{11}+1}{w_{11}+w_{12}+w_{13}+3}(-2 + U^*(w_{11}+1, w_{12}, w_{13}, \ldots))+ \\ & \frac{w_{12}+1}{w_{11}+w_{12}+w_{13}+3}U^*(w_{11}, w_{12}+1, w_{13}, \ldots)+ \\ & \frac{w_{13}+1}{w_{11}+w_{12}+w_{13}+3}(1 + U^*(w_{11}, w_{12}, w_{13}+1, \ldots))\end{aligned}$$

Question 2. Discrete State Filter

Suppose we have a puppy that we need to care for by making sure we give it proper exercise. When the puppy is feeling restless and wants to run around outside, we should let it go outside. However, we cannot directly tell whether the puppy wants to go outside. Instead, we can only observe whether the puppy scratches at the door. We will model this scenario a POMDP with two states, two actions, and two observations. The state, action, and observation spaces are as follows:

$$\mathcal{S} = \{\text{restless}, \text{relaxed}\}$$

$$\mathcal{A} = \{\text{let outside}, \text{ignore}\}$$

$$\mathcal{O} = \{\text{scratch}, \text{no scratch}\}$$

The transition dynamics are:

$$T(\text{restless} \mid \text{restless}, \text{ignore}) = 1.0$$

$$T(\text{restless} \mid \text{restless}, \text{let outside}) = 0.0$$

$$T(\text{restless} \mid \text{relaxed}, \text{let outside}) = 0.0$$

$$T(\text{restless} \mid \text{relaxed}, \text{ignore}) = 0.25$$

When the puppy is restless, it scratches at the door 90 % of the time. When the puppy is relaxed, it still sometimes scratches at the door (because it sees a squirrel or one of its puppy friends outside). It scratches at the door when relaxed 5 % of the time.

a) Explicitly write down the observation model. In other words, what is $O(\text{scratch} \mid \text{restless}, \text{let outside})$, $O(\text{scratch} \mid \text{restless}, \text{ignore})$, $O(\text{scratch} \mid \text{relaxed}, \text{let outside})$, and $O(\text{scratch} \mid \text{relaxed}, \text{ignore})$.

Solution:

$$O(\text{scratch} \mid \text{restless}, \text{let outside}) = 0.9$$

$$O(\text{scratch} \mid \text{restless}, \text{ignore}) = 0.9$$

$$O(\text{scratch} \mid \text{relaxed}, \text{let outside}) = 0.05$$

$$O(\text{scratch} \mid \text{relaxed}, \text{ignore}) = 0.05$$

Note that the action does not affect the observation in this case.

b) Suppose we start with a uniform belief over whether the puppy is restless or relaxed. Let the first element of the belief vector represent the probability that the puppy is restless and the second element represent the probability that the puppy is relaxed. What is our initial belief vector **b**?

Solution:
Since our initial belief is uniform, we believe it is equally likely that the puppy is restless or relaxed. Therefore, our belief vector is $\boxed{\mathbf{b} = [0.5, 0.5]}$.

c) Suppose we do not let the dog out, and we observe that the dog does not scratch at the door. Compute our posterior belief that the dog is restless **up to a proportionality constant**.

Solution:

$$\begin{aligned} b'(\text{restless}) \propto & O(\text{no scratch} \mid \text{restless}, \text{ignore})(T(\text{restless} \mid \text{restless}, \text{ignore})b(\text{restless}) + \\ & T(\text{restless} \mid \text{relaxed}, \text{ignore})b(\text{relaxed})) \\ = & 0.1(1.0(0.5) + 0.25(0.5)) \\ = & \boxed{0.0625} \end{aligned}$$

d) Compute our posterior belief that the dog is relaxed **up to a proportionality constant**.

Solution:

$$\begin{aligned} b'(\text{relaxed}) \propto & O(\text{no scratch} \mid \text{relaxed}, \text{ignore})(T(\text{relaxed} \mid \text{restless}, \text{ignore})b(\text{restless}) + \\ & T(\text{relaxed} \mid \text{relaxed}, \text{ignore})b(\text{relaxed})) \\ = & 0.95(0.0(0.5) + 0.75(0.5)) \\ = & \boxed{0.3562} \end{aligned}$$

e) Based on the results from part (c) and part (d), compute our new belief vector $\mathbf{b}'$.

Solution:
We need to normalize the results from part (b) and part (c) so that our new belief vector sums to 1. We compute

$$\mathbf{b}' = [0.0625, 0.3562]/(0.0625 + 0.3562)$$

$$\boxed{\mathbf{b}' = [0.1493, 0.8507]}$$