Reinforcement Learning Computational Problems

Based on Quiz Review Notes

2024

Given the action preferences $\{p_1, p_2, p_3\} = \{1.2, 0.5, 0.1\}$, the probability of selecting each action using the softmax policy is calculated as follows:

The softmax policy assigns the probability for action i as:

$$\pi(a_i) = \frac{e^{p_i}}{\sum_{j=1}^{3} e^{p_j}}$$

For the given preferences:

$$\pi(a_1) = \frac{e^{1.2}}{e^{1.2} + e^{0.5} + e^{0.1}} = \frac{e^{1.2}}{e^{1.2} + e^{0.5} + e^{0.1}}$$

$$\pi(a_2) = \frac{e^{0.5}}{e^{1.2} + e^{0.5} + e^{0.1}}$$

$$\pi(a_3) = \frac{e^{0.1}}{e^{1.2} + e^{0.5} + e^{0.1}}$$

Now calculating the values of e^{p_i}:

$$e^{1.2} \approx 3.3201, \quad e^{0.5} \approx 1.6487, \quad e^{0.1} \approx 1.1052$$

The denominator becomes:

$$3.3201 + 1.6487 + 1.1052 = 6.074$$

Thus, the probabilities are:

$$\pi(a_1) \approx \frac{3.3201}{6.074} \approx 0.5466$$

$$\pi(a_2) \approx \frac{1.6487}{6.074} \approx 0.2714$$

$$\pi(a_3) \approx \frac{1.1052}{6.074} \approx 0.1820$$

Therefore, the probabilities of selecting actions a_1, a_2, and a_3 are approximately 0.5466, 0.2714, and 0.1820, respectively.

Problem 1: Policy Gradient Estimation

Given the following policy $\pi_\theta(s, a)$ for state s and action a:

$$\pi_\theta(s, a) = \frac{e^{\phi(s,a)^T \theta}}{\sum_{a' \in A} e^{\phi(s,a')^T \theta}},$$

where $\phi(s, a)$ is the feature vector for state-action pair (s, a), and θ is the parameter vector.

Given:

$$\phi(s, a_1) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \phi(s, a_2) = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \quad \theta = \begin{bmatrix} 0.3 \\ 0.7 \end{bmatrix},$$

compute the probability of taking action a_1 in state s, i.e., $\pi_\theta(s, a_1)$.

Solution

First, calculate $\phi(s, a_1)^T \theta$ and $\phi(s, a_2)^T \theta$:

$$\phi(s, a_1)^T \theta = 1 \cdot 0.3 + 2 \cdot 0.7 = 1.7,$$

$$\phi(s, a_2)^T \theta = 0.5 \cdot 0.3 + 1 \cdot 0.7 = 0.85.$$

Next, compute the action probabilities:

$$\pi_\theta(s, a_1) = \frac{e^{1.7}}{e^{1.7} + e^{0.85}} = \frac{e^{1.7}}{e^{1.7} + e^{0.85}}.$$

Using a calculator:

$$\pi_\theta(s, a_1) = \frac{5.4739}{5.4739 + 2.3396} = \frac{5.4739}{7.8135} \approx 0.7005.$$

Thus, $\pi_\theta(s, a_1) \approx 0.7005$.

Problem 2: Gaussian Policy Log-Probability

In a Gaussian policy, the action a is sampled from a distribution $\mathcal{N}(\mu_\theta(s), \sigma_\theta(s)^2)$.

Given:

$$\mu_\theta(s) = 2, \quad \sigma_\theta(s) = 0.5, \quad a = 3,$$

compute the log-probability $\log \pi_\theta(a|s)$.

Solution

The log-probability for a Gaussian distribution is given by:

$$\log \pi_\theta(a|s) = -\frac{1}{2} \log(2\pi\sigma_\theta(s)^2) - \frac{(a - \mu_\theta(s))^2}{2\sigma_\theta(s)^2}.$$

Substitute the given values:

$$\log \pi_\theta(a|s) = -\frac{1}{2}\log(2\pi(0.5)^2) - \frac{(3-2)^2}{2(0.5)^2}.$$

Calculate each term:

$$-\frac{1}{2}\log(2\pi(0.25)) = -\frac{1}{2}\log(1.5708) \approx -0.2857,$$

$$\frac{(3-2)^2}{2(0.5)^2} = \frac{1}{2 \cdot 0.25} = 2.$$

Thus, the log-probability is:

$$\log \pi_\theta(a|s) \approx -0.2857 - 2 = -2.2857.$$

Problem 3: Expected Reward in One-Step MDP

In a one-step MDP, the agent takes an action a, receives a reward $R(s,a)$, and terminates. Given the policy:

$$\pi(a|s) = \begin{cases} 0.6 & \text{if } a = a_1, \\ 0.4 & \text{if } a = a_2, \end{cases}$$

and the rewards:

$$R(s,a_1) = 10, \quad R(s,a_2) = 5,$$

compute the expected reward $V^\pi(s)$.

Solution

The expected reward is:

$$V^\pi(s) = \sum_a \pi(a|s) R(s,a).$$

Substitute the values:

$$V^\pi(s) = 0.6 \cdot 10 + 0.4 \cdot 5 = 6 + 2 = 8.$$

Thus, the expected reward $V^\pi(s) = 8$.

Problem 4: Gradient with Respect to Policy Parameters

Consider a softmax policy with two actions. The probabilities of actions a_1 and a_2 are given by:

$$\pi(a_1|s;\theta) = \frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_2}}, \quad \pi(a_2|s;\theta) = \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2}}.$$

Compute the gradient $\nabla_{\theta_1} \log \pi(a_1|s;\theta)$.

Solution

The gradient of the log-probability is given by:

$$\nabla_{\theta_1} \log \pi(a_1|s;\theta) = 1 - \pi(a_1|s;\theta).$$

Substitute $\pi(a_1|s;\theta)$:

$$\nabla_{\theta_1} \log \pi(a_1|s;\theta) = 1 - \frac{e^{\theta_1}}{e^{\theta_1} + e^{\theta_2}} = \frac{e^{\theta_2}}{e^{\theta_1} + e^{\theta_2}}.$$

Thus, $\nabla_{\theta_1} \log \pi(a_1|s;\theta) = \pi(a_2|s;\theta)$.

Problem 5: Monte-Carlo Policy Gradient

In a Monte-Carlo policy gradient method, the return G_t at time step t is the cumulative discounted reward:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k}),$$

where $\gamma = 0.9$. Given rewards $R(s_t, a_t) = 5$, $R(s_{t+1}, a_{t+1}) = 3$, $R(s_{t+2}, a_{t+2}) = 2$, compute G_t.

Solution

Compute G_t using the given rewards:

$$G_t = 5 + 0.9 \cdot 3 + 0.9^2 \cdot 2 = 5 + 2.7 + 1.62 = 9.32.$$

Thus, $G_t = 9.32$.