An Introduction to Reinforcement Learning

Dongyan Lin

Imbizo 2023

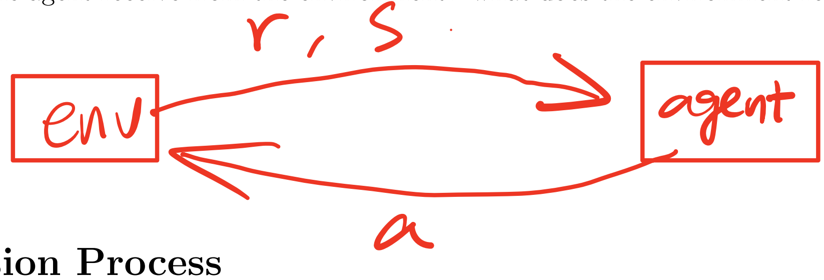
1 Basic Concepts

Answer in your own words:

- What is reinforcement learning? How is it different from supervised learning and unsupervised learning?

  learning through trial & error | learning through labeled data (i.e. target) | learning through unlabeled data

- Draw a diagram demonstrating the interaction between the environment and the agent. Specify on your diagram: what does the agent receive from the environment? what does the environment receive from the agent?



2 Markov Decision Process

- What does it mean for a state S_t to satisfy the Markov property? Write down both the mathematical definition and the verbal meaning.

  $$Pr[S_{t+1} | S_t] = Pr[S_{t+1} | S_1, S_2, \dots S_{t-1}, S_t]$$

  Given the present, the future does not depend on the past.

- Below are the variables that define a Markov Decision Process, What does each of them denote?
  - $\mathcal{S}$ set of states $= \langle s_1, s_2, \dots \rangle$
  - $\mathcal{A}$ set of actions $= \langle a_1, a_2, \dots \rangle$
  - $\mathcal{P}$ transition probability matrix
  - $\mathcal{R}$ reward function
  - γ discount factor
- What is the mathematical definition of the return?
  - $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
- What is the mathematical definition of the value functions?
  - state-value $v(s) = \mathbb{E}[G_t \mid S_t = s]$
  - state-action-value $q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$
- What is the mathematical definition of the policy?
  - $\pi[a|s] = pr(A_t = a \mid S_t = s)$
- (Optional) Note: not all RL problems are Markovian (in fact, the Markov property is a very hard constraint to satisfy). Can you think of an example that involves learning through trial-and-error, but does not satisfy the Markov property?

  In short, anything that requires memory/knowledge about the past, in addition to the current state.

  Eg. a working memory decision making task, stock market, escape room ...

  (Note: chess is Markovian! Because the current configuration of the chess board is informative enough to make a decision)

3 The Environment

this means number of elements in a set.

- The reward function $R(s)$ is defined as the reward R the agent receives upon leaving the state s. It is usually represented as a 1-dimensional vector. What is the shape (i.e., length of each dimension) of the vector $R(s)$?

  $|S| \times 1$ i.e., number of states.

- The transition function $P(s, s', a)$ is defined as the probability of moving from state s to s' when the agent takes action a. It is usually represented by a 3-dimensional tensor $P(s, s', a)$. What is the shape of the tensor $P(s, s', a)$?

  $|S| \times |S| \times |\mathcal{A}|$

- (Optional) Sometimes the reward R that the agent receives upon leaving the state s also depends on the action it takes, a. In this case, the reward function becomes a 2-dimensional matrix, $R(s, a)$. What is the shape of the matrix $R(s, a)$?

  $|S| \times |\mathcal{A}|$

4 The Agent

- Explain in your own words: what are the differences between model-free and model-based reinforcement learning? Hint: it might help to first answer the question: what is a "model"?

  model = transition probability matrix + reward function.
  MF: Solve RL problem w/ value & policy.
  MB: Solve RL problem w/ model. more flexible, but need more compute.

- Greedy and ϵ-greedy policy
  - What does it mean to follow a greedy policy?
    * $a_t = \arg\max_{a \in A} q(s,a)$

    What is the mathematical definition of a greedy policy?
    * $\pi(s) = \arg\max_{a \in A} q(s,a)$
  - What does it mean to follow a ϵ-greedy policy?
    * $a_t = \begin{cases} \arg\max_{a \in A} q(s,a) & \text{with } (1-\varepsilon) \text{ probability} \\ \text{random action} & \text{with } \varepsilon \text{ probability} \end{cases}$

    What is the mathematical definition of a ϵ-greedy policy?
    * $\pi[a|s] = \begin{cases} \frac{\varepsilon}{m} + (1-\varepsilon) & \text{if } a = \arg\max_{a \in A} q(s,a) \\ \varepsilon/m & \text{otherwise} \end{cases}$

    proof: see next page

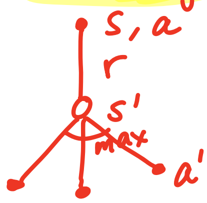
- Describe the 3 ways to compute value of the current state, $v_\pi(s)$.
  - Brute Force: all possible transitions and actions considered
  - Sampling (eg. MC): average the returns of sampled trajectories
  - Bootstrapping (eg. TD): get a reward, then add to the value estimate of the next state

- (Optional) Derive the incremental update rule for learning value, $V_t = V_{t-1} + \alpha(G(\tau_N) - V_{t-1})$, from the equation for calculating value by sampling trajectories, $V_t = \frac{1}{N}\sum_{i=1}^{N} G(\tau_i)$.

  see next page.

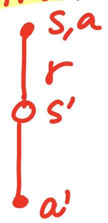
- Describe the learning rule and draw the one-step look-ahead diagram for Q-learning and SARSA.

Q-learning:



$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$

SARSA:



$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$

$$V_t = \frac{1}{N}\sum_{i=1}^{N} G(\tau_i)$$
$$= \frac{1}{N}\left[G(\tau_N) + \sum_{i=1}^{N-1} G(\tau_i)\right]$$
$$= \frac{1}{N}\left[G(\tau_N) + (N-1)\underbrace{\boxed{\frac{1}{N-1}\sum_{i=1}^{N-1} G(\tau_i)}}\right]$$
$$= \frac{1}{N}\left[G(\tau_N) + (N-1)\ \boxed{V_{t-1}}\ \right]$$
$$= \frac{1}{N}\left[G(\tau_N) + N V_{t-1} - V_{t-1}\right]$$
$$= \frac{1}{N}G(\tau_N) + \cancel{\frac{1}{N}}\cdot \cancel{N} V_{t-1} - \frac{1}{N}V_{t-1}$$
$$= V_{t-1} - \frac{1}{N}\left[V_{t-1} - G(\tau_N)\right]$$
$$= V_{t-1} - \alpha\left[V_{t-1} - G(\tau_N)\right] \quad \text{if we let } \alpha = \frac{1}{N}.$$

<u>Assume:</u>
sample trajectory τ_N at time t
sample trajectory τ_{N-1} at time $t-1$
⋮

$$a = \begin{cases} \underset{a \in A}{\operatorname{argmax}}\ q(s,a) & \text{with } (1-\varepsilon) \text{ probability} \\ \text{random action} & \text{with } \varepsilon \text{ probability} \end{cases}$$

Say, there are m possible actions.

One of them is the greedy action, i.e. $\underset{a \in A}{\operatorname{argmax}}\ q(s,a)$

then, when we randomly sample an action, there is still $\frac{\varepsilon}{m}$ possibility that we'll sample the greedy action.

$$\Rightarrow Pr\left(\underset{a \in A}{\operatorname{argmax}}\ q(s,a)\right) = \frac{\varepsilon}{m} + (1-\varepsilon)$$