Aula 1 - Introduction to Reinforcement Learning

Gabriel Valentim

March 2023

1 Introduction

Aprendizado por reforço (*Reinforcement Learning*, ou RL) refere-se ao conjunto de técnicas para o treinamento de agentes autônomos através de incentivos que chamamos de recompensa. Um agente pode ser qualquer entidade que se encontra em um determinado ambiente, com o intuito de realizar alguma tarefa. Nesse sentido, o agente interage com o ambiente e isso caracteriza a essência do sistema que é o foco do estudo do *Reinforcement Learning*, tal como visto na Figura 1.

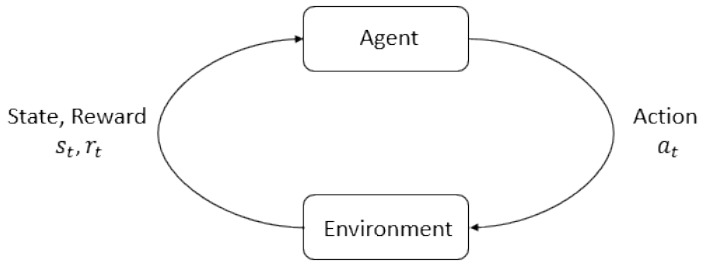

Figure 1: Interação agente-ambiente.

Na Figura 1 é possível descrever o comportamento e o andamento do processo com a descrição de algumas variáveis que são pertinentes à condição do Agente: a ação a_t a ser tomada no tempo $t \in \{0, 1, 2, \cdots\}$, o estado s_t do tempo t, o próximo estado s_{t+1}, alcançado após a tomada da ação a_t e a recompensa r_t devido a troca de estado.

A exemplo disso, podemos pensar num robô aspirador de pó que traça trajetórias com o intuito de limpar uma sala. Nesse contexto, o agente é o robô e o ambiente é sala. Nosso sistema buscará resolver o problema central do *Reinforcement Learning*, que é maximizar a recompensa acumulada, que nesse caso, é fazer com que o robô limpe a sala no menor tempo possível, traçando trajetórias que são optimizadas para que isso ocorra.

Dessa forma, vamos definir alguns aspectos importantes e necessários para que possamos descrever e entender os sistemas de *Reinforcement Learning* de

uma maneira mais consistente e estruturada, sobretudo, fazendo uso das ferramentas matemáticas para descrever como funcionam alguns conceitos e abordagens.

2 *Policy*

A *policy* é a regra utilizada pelo agente para decidir quais ações tomar. Uma *policy* pode ser determinística, na qual as ações são uma função do estado do ambiente, ou estocástica, onde as ações são escolhidas a partir de uma distribuição estocástica condicionada ao estado do ambiente. Neste trabalho vamos estudar *policies* estocásticas. As *policies* estocásticas são usualmente denotadas por π:

$$a_t \sim \pi(\cdot|s_t). \tag{1}$$

A Equação 1 indica que a ação a_t a ser tomada pelo agente é obtida como uma amostra aleatória da distribuição condicional $\pi(\cdot|s_t)$. No caso limite em que a distribuição condicional converge para uma distribuição delta de Dirac, a *policy* torna-se determinística.

É comum que a *policy* seja mencionada de forma intercambiável com o agente, pelo fato da policy poder ser interpretada como o cérebro do agente. Comumente lidamos com *policies* parametrizadas, onde denotamos os parâmetros de tal *policy* por θ ou ϕ e, em seguida escrevemos isso como um subscrito no símbolo da *policy* (Equação 2):

$$a_t \sim \pi_\theta(\cdot|s_t). \tag{2}$$

3 Trajetória

Uma trajetória é uma sequência de estados e ações de um agente que atua em um determinado ambiente, para realizar uma determinada tarefa. Lembre-se que utilizamos as notações de s_t e a_t para representar, respectivamente, o estado e a ação no tempo t.

$$\tau = (s_0, a_0, s_1, a_1, ...) \tag{3}$$

O primeiro estado é aleatoriamente inserido seguindo uma distribuição de estado inicial denotada por ρ_0:

$$s_0 \sim \rho_0(\cdot).$$

A transição de estado do ambiente de s_t para s_{t+1} depende de modo probabilístico do estado atual do ambiente (s_t) e da ação tomada pelo agente (a_t), conforme ilustrado na Equação 4:

$$s_{t+1} \sim P(\cdot|s_t, a_t). \tag{4}$$

4 Recompensa e Retorno

A função de recompensa $\mathcal{R}$ é criticamente importante para o *Reinforcement Learning*. Isso depende do estado atual (s_t), da ação tomada (a_t), e o próximo estado (s_{t+1}):

$$r_t = \mathcal{R}(s_t, a_t, s_{t+1}). \tag{5}$$

Em conceitos mais palpáveis, a função de recompensa nos traz a influência da transição entre os estados (tomada a ação a_t) de forma quantitativa. Sobretudo, isso tem forte relação com a tarefa que deve ser realizada pelo agente, naquele ambiente. Nesse sentido, vale ressaltar que existe alguns tipos de cálculo do retorno.

Dada uma sequência de recompensas $(r_0, r_1, \cdots, r_T)$ em uma trajetória τ, podemos definir o retorno completo $R(\tau)$ sobre a trajetória de diferentes formas, conforme visto abaixo. O objetivo do *Reinforcement Learning* é a construção de uma *policy* que maximize o retorno médio sobre todas as trajetórias.

Um tipo de retorno é o *finite-horizon undiscounted return*, que é a soma das recompensas obtidas (fixas) para cada passo, sobre toda a trajetória:

$$R(\tau) = \sum_{t=0}^{T} r_t. \tag{6}$$

Um outro tipo de retorno é o *infinite-horizon discounted return*, que é a soma das recompensas obtidas para cada passo, multiplicado por um fator de desconto $\gamma \in (0, 1)$:

$$R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t. \tag{7}$$

Note que $\gamma \in (0, 1)$ é muito conveniente pois, além de garantir a convergência da soma, demonstra que a recompensa maior será atribuída para ação tomada no menor tempo, o que faz sentido para o treinamento do agente.

Por fim, é possível utilizar também o *reward to-go*. Nesse caso, levaremos em conta somente as recompensas acumuladas do instante t em diante, ou seja, até o fim da trajetória.

$$G_t = \sum_{t'=t}^{T} \mathcal{R}(s_{t'}, a_{t'}, s_{t'+1}), \tag{8}$$

O que faz todo sentido, haja vista o fato de que a atualização da *policy* deve ser tomada de acordo com a recompensa tomada do instante t atual até o fim da trajetória. Mais adiante vamos utilizar essa função de recompensa para a construção do algoritmo *REINFORCE* e vamos mostrar que a utilização dela pode ser feita no teorema do *policy gradient* sem problemas. Por fim, vamos definir o retorno esperado como $J(\pi)$:

$$J(\pi) = E_{\tau \sim \pi}[R(\tau)] = \int_{\tau} P(\tau|\pi) R(\tau) d\tau \tag{9}$$