Aprendizagem por Reforço

Wuerike Henrique da Silva Cavalheiro

gym_taxi.py

Sumário

- Introdução
- Processo de decisão de Markov
- Retorno esperado
- Políticas
- Funções de valor
- Política ótima
- Q-Learning
- Conceitos complementares

Introdução

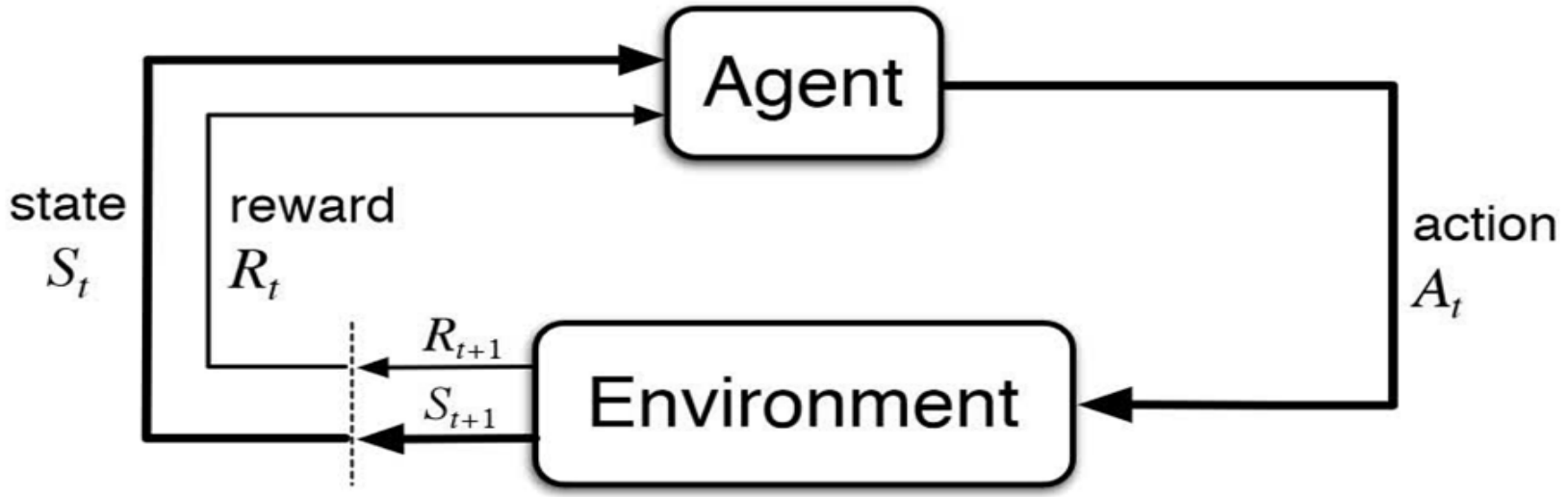
- Aprendizagem supervisionada
  - Generalização de novos dados partindo de exemplos
- Aprendizagem não supervisionada
  - Identificação de padrões em dados não rotulados
- Aprendizagem por reforço
  - Aprende como se comportar através da iteração com ambiente
  - Um agente interage com o ambiente, buscando tomar ações que tragam a maior recompensa

Processo de decisão de Markov

- Propriedade de Markov
  - O futuro é independente do passado para um dado presente
- Elementos de um Processo de Decisão de Markov
  - Agente
  - Ambiente
  - Estado – S
  - Ação – A
  - Recompensa – R
- O objetivo do agente é maximizar as recompensas acumuladas durante sua trajetória

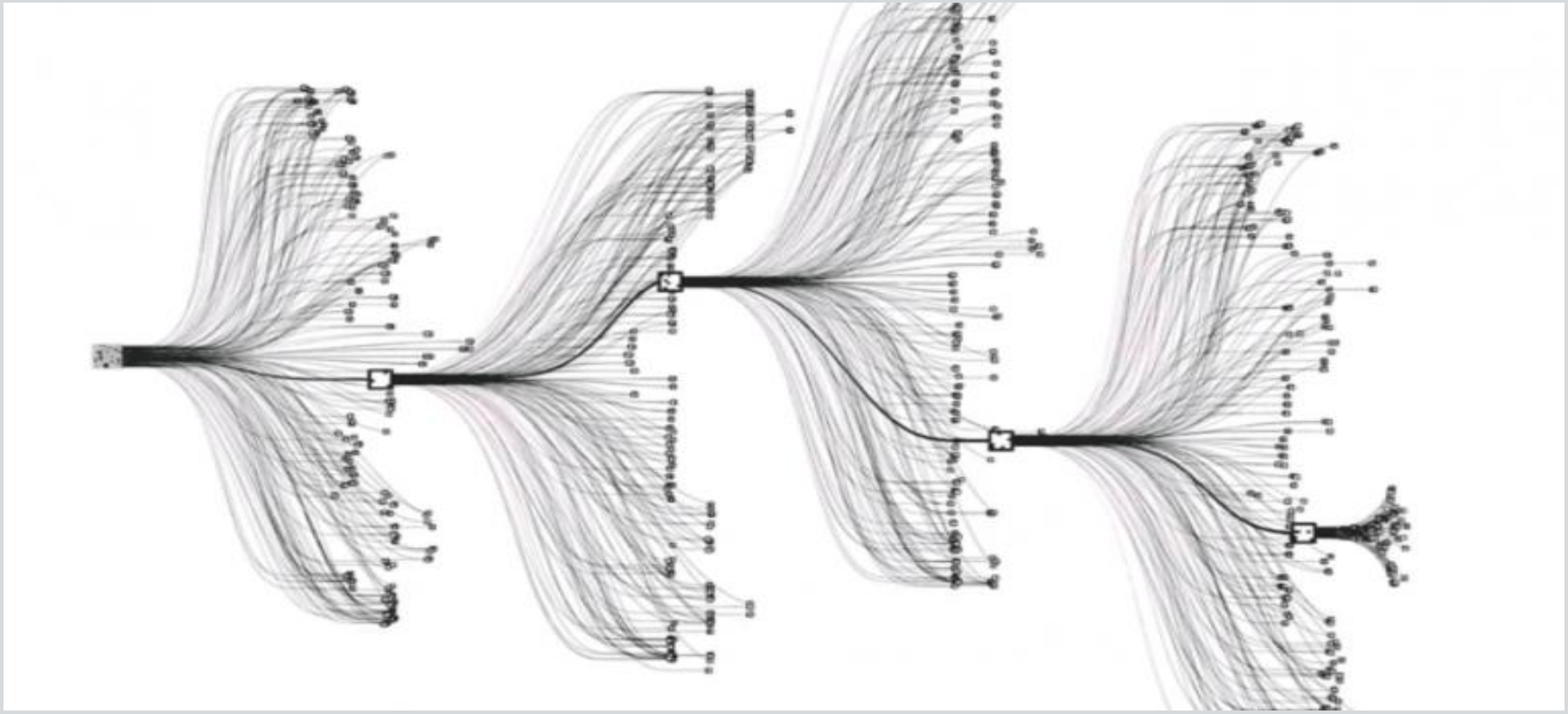
Processo de decisão de Markov

- Fluxo de um Processo de Decisão de Markov
  - No tempo t, o ambiente esta no estado S
  - O agente observa o ambiente e realiza a ação A_t
  - O ambiente muda para S_{t+1} e retorna a recompensa R_{t+1}
  - O processo é executado novamente para $t + 1$



Processo de decisão de Markov

A realização de uma ação leva a um estado onde diversas outras ações podem ser tomadas



Retorno esperado

- É o somatório das recompensas esperadas no futuro
  - $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots + R_T$
- Retorno Descontado
  - $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$
- Cenário para $R = 1$ constante e $\gamma < 1$
  - $G_t = \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$
- O objetivo do agente é maximizar o retorno descontado esperado durante sua trajetória

Políticas

- Nos fala sobre o comportamento do agente
  - Qual a probabilidade de um agente selecionar um ação específica quando estiver em um determinado estado?

- A politica π é uma função que retorna a probabilidade da escolha de uma ação em um estado.

- Se um agente segue a política π no instante t, então $\pi(a|s)$ é a probabilidade de $A_t = a$ se $S_t = s$.

Funções de valor

- Estima o quão bom é para o agente tomar uma determinada ação estando num certo estado

- O valor da ação a no estado s seguindo a política π equivale ao retorno esperado descontado dados s e a

$$q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

- A função de valor q_π é normalmente chamada de Q-function e o seu valor é chamado de Q-value

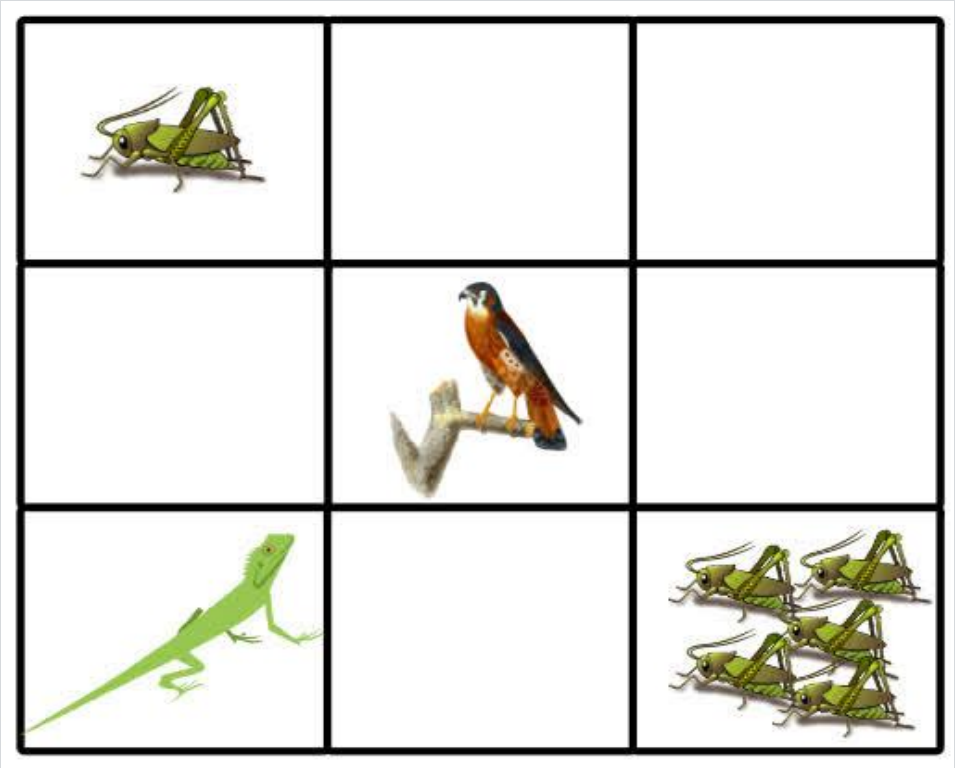
Política ótima

- Dadas as políticas π e π'
  - $\pi \geq \pi'$ apenas se $q_\pi(s,a) \geq q_{\pi'}(s,a)$ para todo $s \in S$ e $a \in A(s)$
- A política ótima tem uma função valor ótima
  - $q_*(s,a) = \max_\pi q_\pi(s,a)$
- Equação de Bellman
  - $q_*(s,a) = E\left[R_{t+1} + \gamma \max_{a'} q_*(s',a')\right]$

Q-Learning

Resolução do MDP aplicando a Equação de Bellman

$$q_{new}(s,a) = (1-\alpha)q_{old}(s,a) + \alpha\left[R_{t+1} + \gamma \max_{a'} q_*(s',a')\right]$$

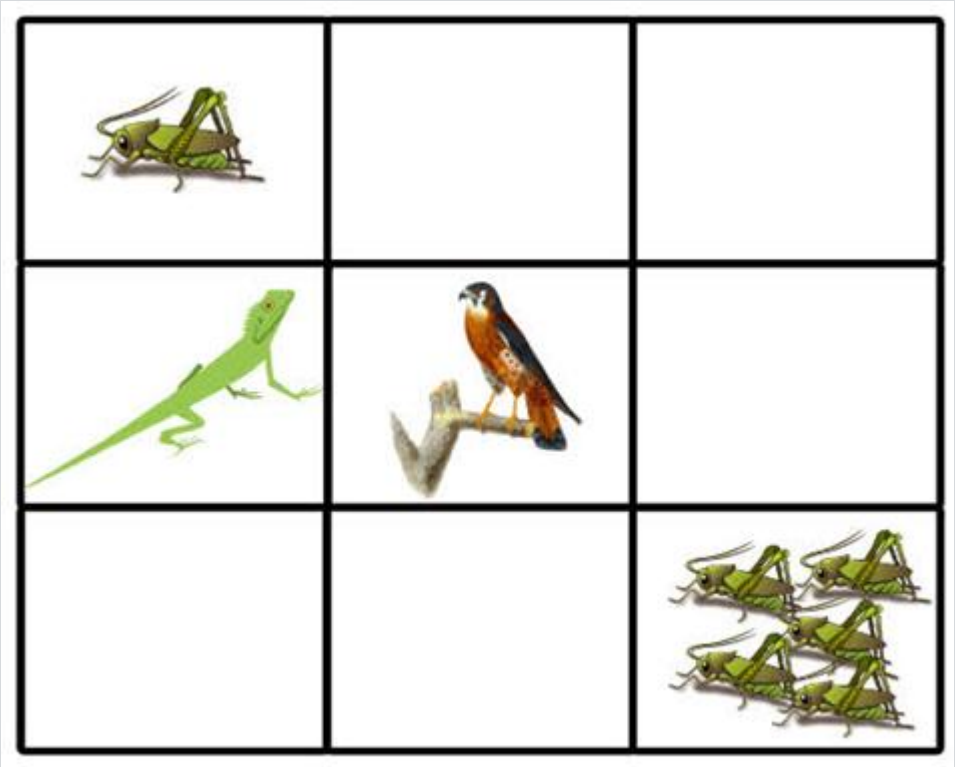


		Actions			
		Left	Right	Up	Down
States	1 cricket	0	0	0	0
	Empty 1	0	0	0	0
	Empty 2	0	0	0	0
	Empty 3	0	0	0	0
	Bird	0	0	0	0
	Empty 4	0	0	0	0
	Empty 5	0	0	0	0
	Empty 6	0	0	0	0
	5 crickets	0	0	0	0

Q-Learning

Resolução do MDP aplicando a Equação de Bellman

$$q_{new}(s,a) = (1-\alpha)q_{old}(s,a) + \alpha\left[R_{t+1} + \gamma \max_{a'} q_*(s',a')\right]$$

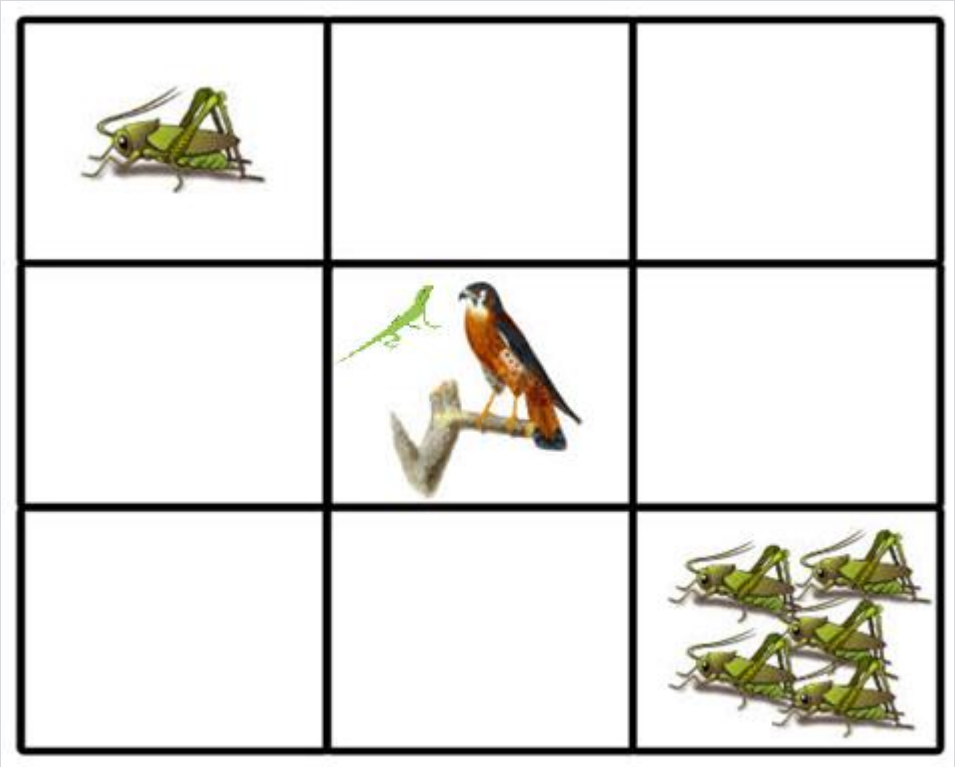


		Actions			
		Left	Right	Up	Down
States	1 cricket	0	0	0	0
	Empty 1	0	0	0	0
	Empty 2	0	0	0	0
	Empty 3	0	0	0	0
	Bird	0	0	0	0
	Empty 4	0	0	0	0
	Empty 5	0	0	-1	0
	Empty 6	0	0	0	0
	5 crickets	0	0	0	0

Q-Learning

Resolução do MDP aplicando a Equação de Bellman

$$q_{new}(s,a) = (1-\alpha)q_{old}(s,a) + \alpha\left[R_{t+1} + \gamma \max_{a'} q_*(s',a')\right]$$



		Actions			
		Left	Right	Up	Down
States	1 cricket	0	0	0	0
	Empty 1	0	0	0	0
	Empty 2	0	0	0	0
	Empty 3	0	-10	0	0
	Bird	0	0	0	0
	Empty 4	0	0	0	0
	Empty 5	0	0	-1	0
	Empty 6	0	0	0	0
	5 crickets	0	0	0	0

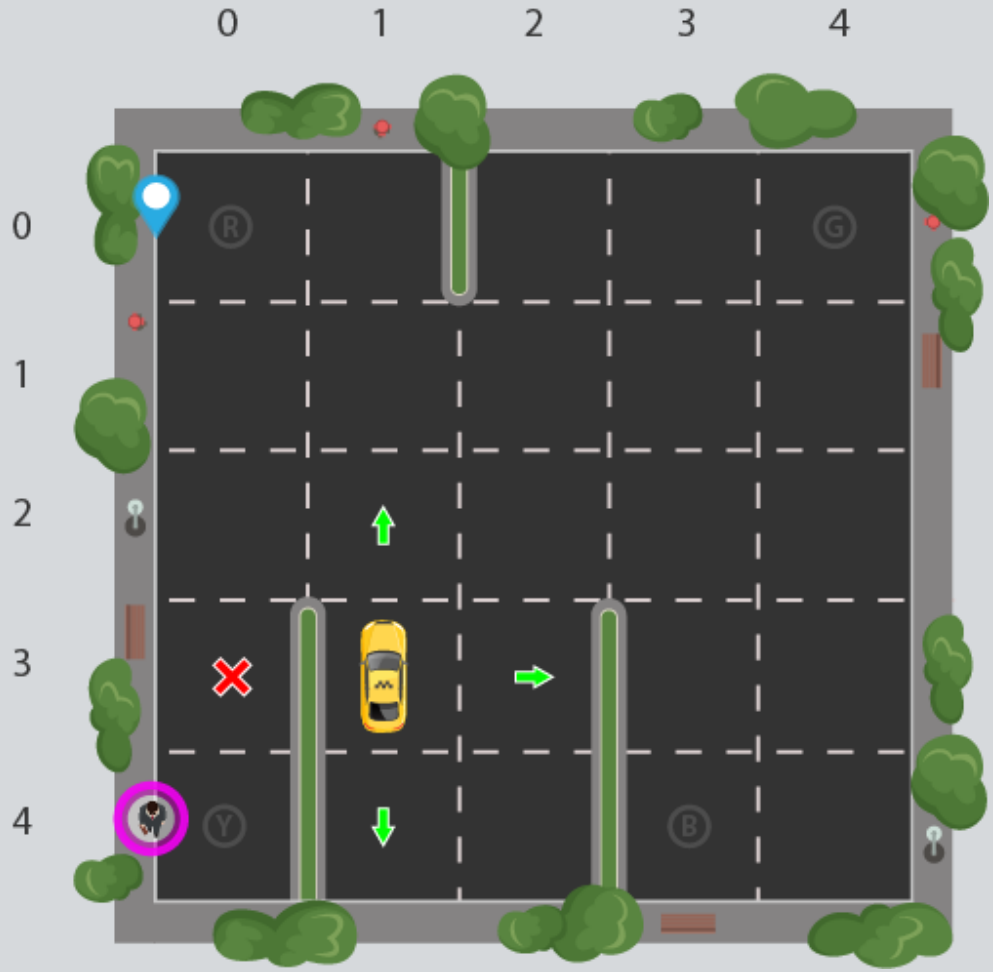
Q-Learning

Ações:

- Norte
- Sul
- Leste
- Oeste
- Embarque
- Desembarque

Recompensas:

- -1 por ação
- -10 por ação impossível
- +20 ao finalizar



Conceitos complementares

- Delayed reward
- Episodic vs Continuing Tasks
- Models
- State-Value Function
- Exploration vs exploitation tradeoff
- Learning Rate

Contato

E-mail: wuerike.henri@gmail.com

Telefone: (47) 99648-9327

UFSC
Joinville