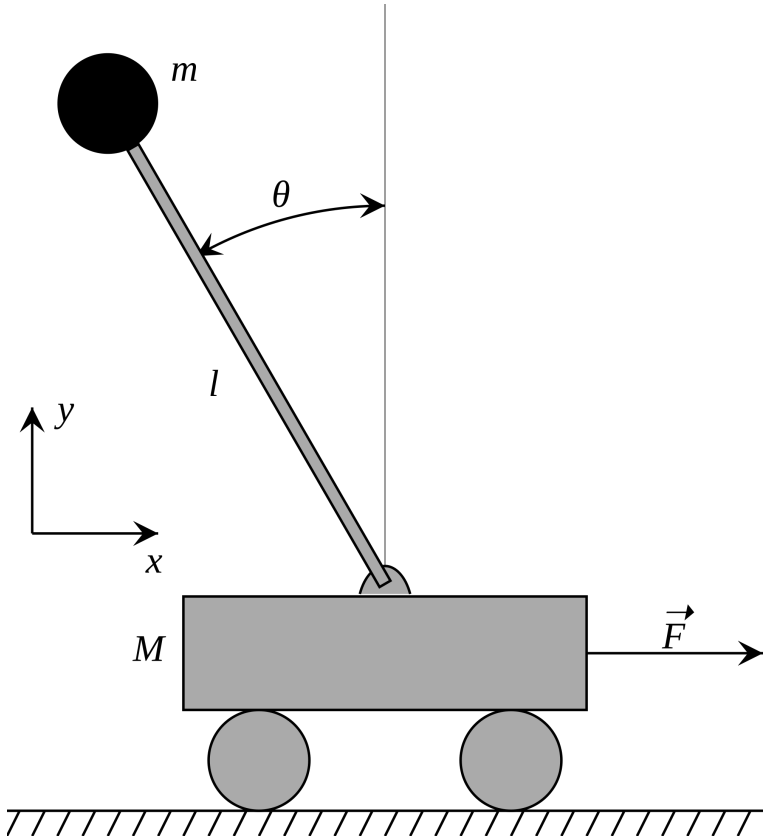
Planejamento Probabilístico e Aprendizado por Reforço

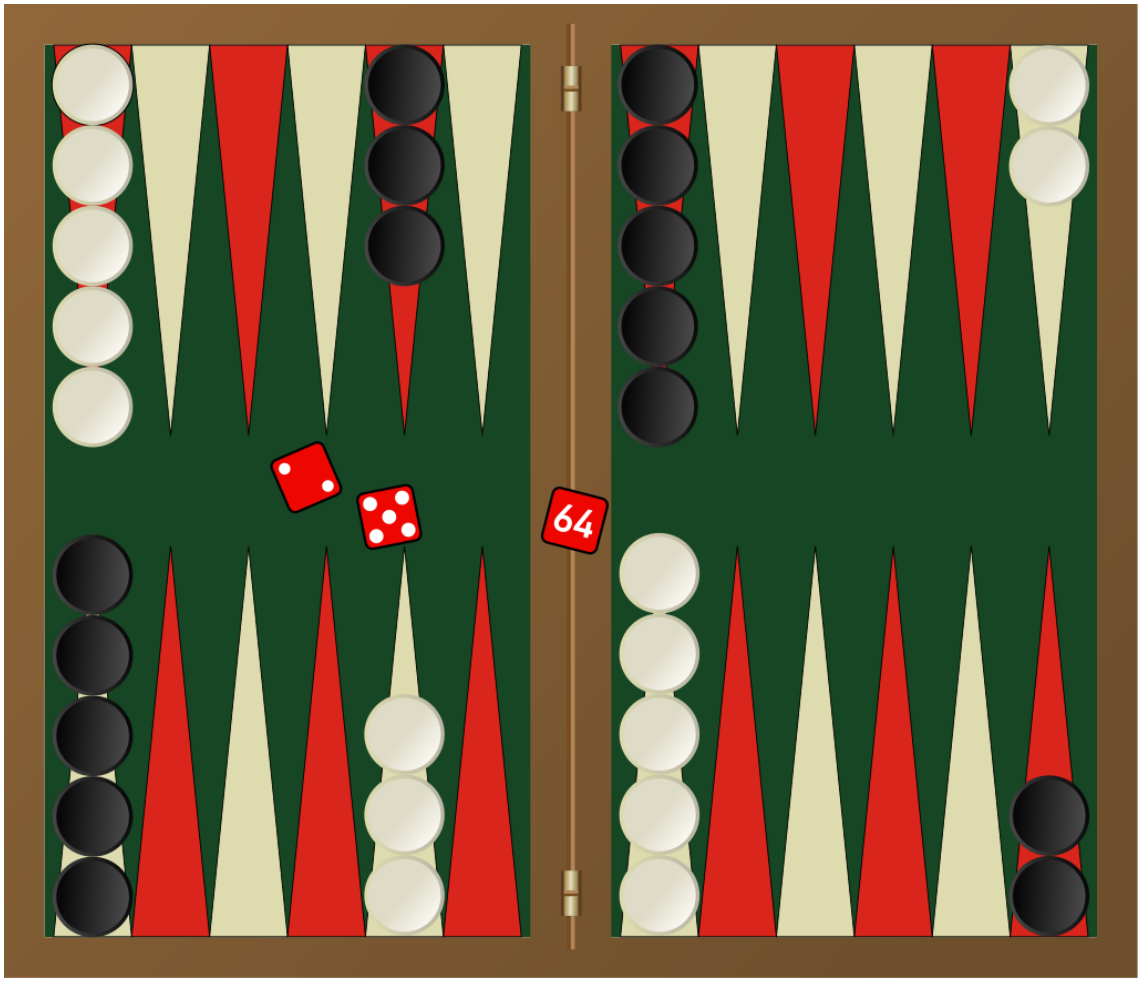
Valdinei Freire

(EACH - USP)

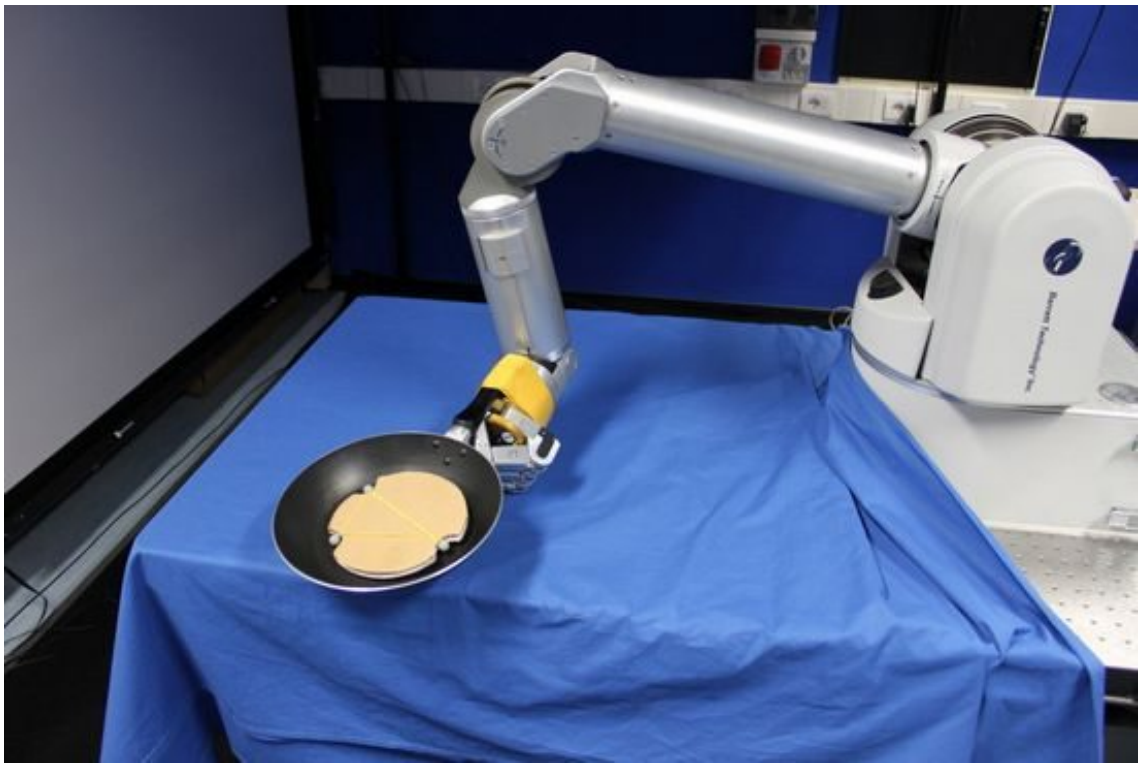
Cart-Pole



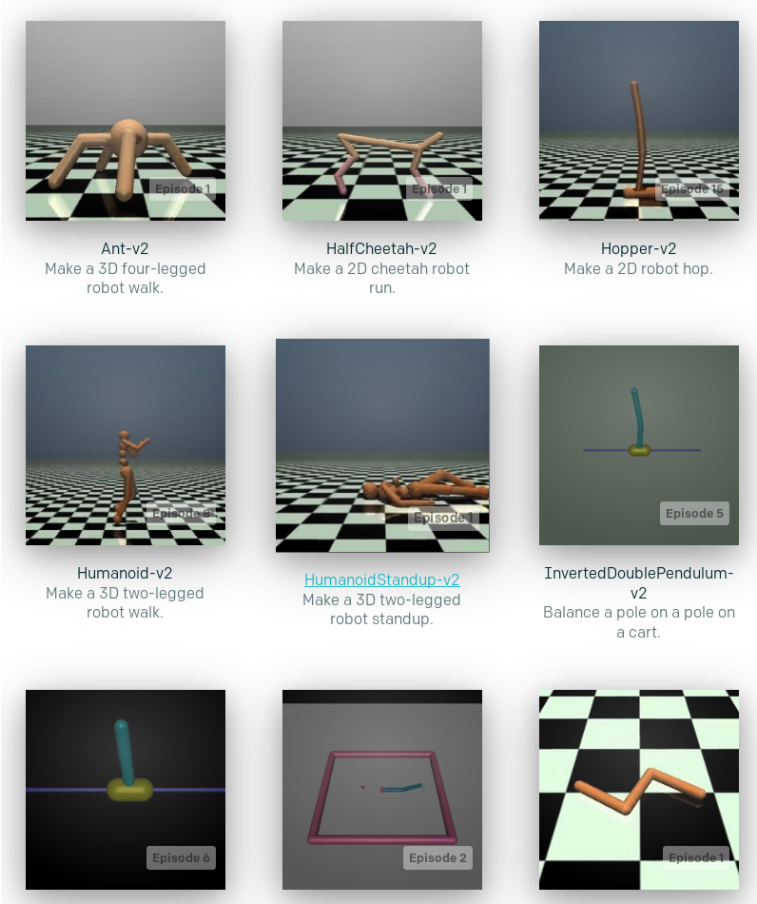
Gammão



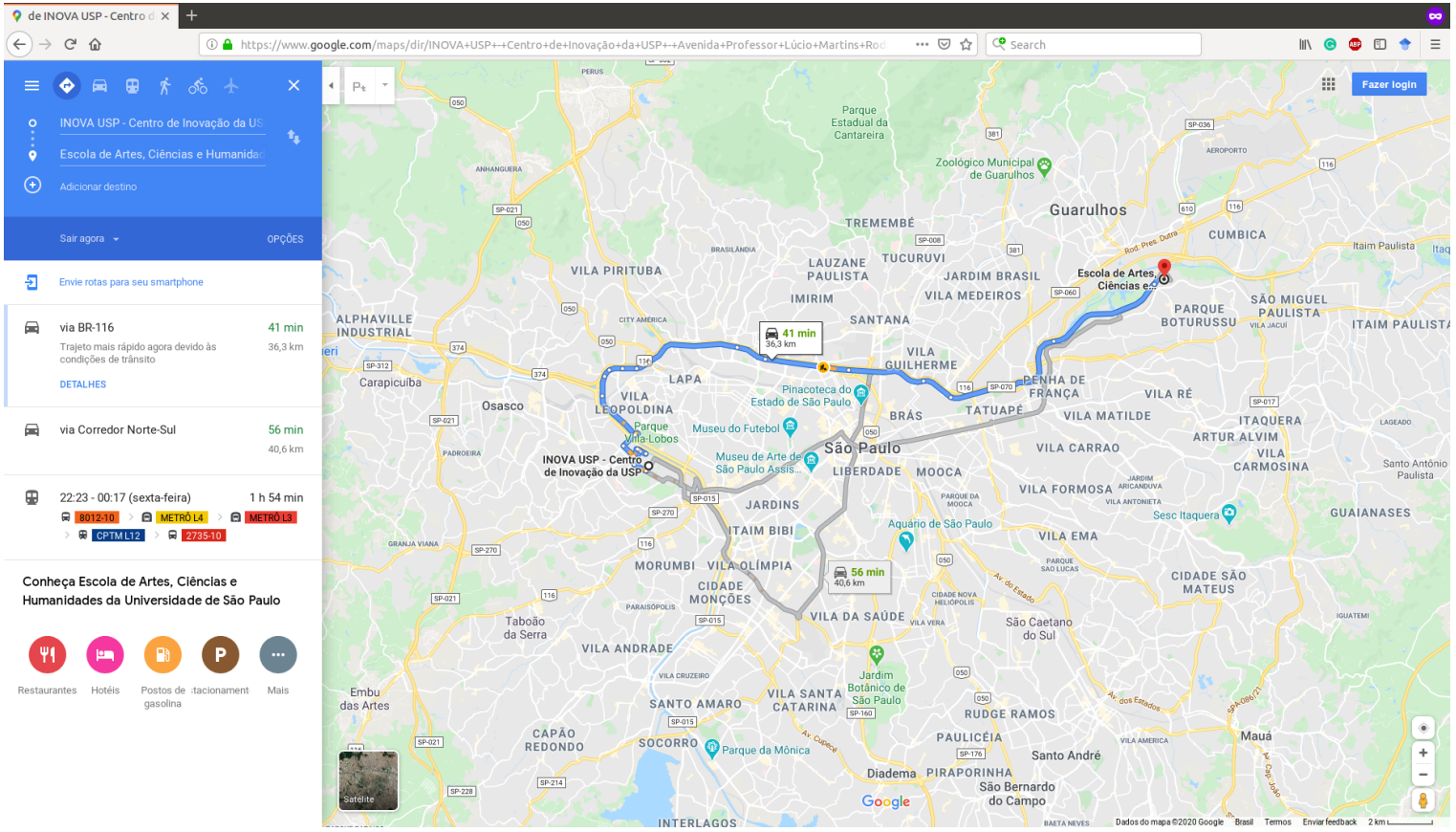
Robô



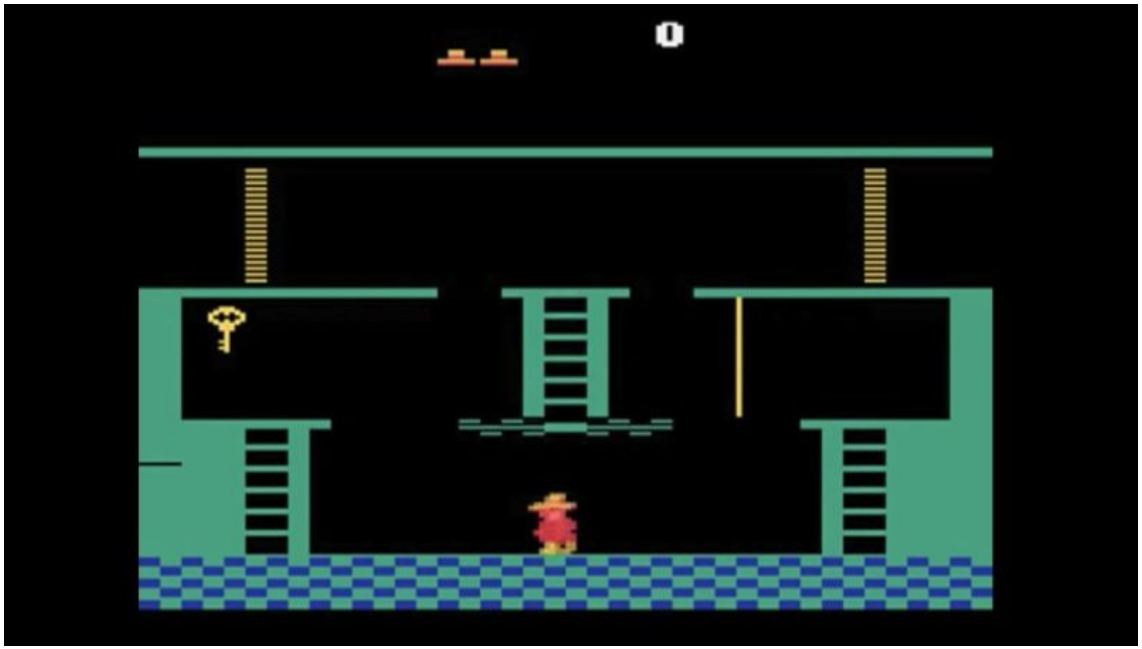
OpenAI GYM


Episode 1
Ant-v2
Make a 3D four-legged robot walk.
Episode 1
HalfCheetah-v2
Make a 2D cheetah robot run.
Hopper-v2
Make a 2D robot hop.
Humanoid-v2
Make a 3D two-legged robot walk.
Episode 1
HumanoidStandup-v2
Make a 3D two-legged robot standup.
Episode 5
InvertedDoublePendulum-v2
Balance a pole on a pole on a cart.
Episode 6
Episode 2
Episode 1

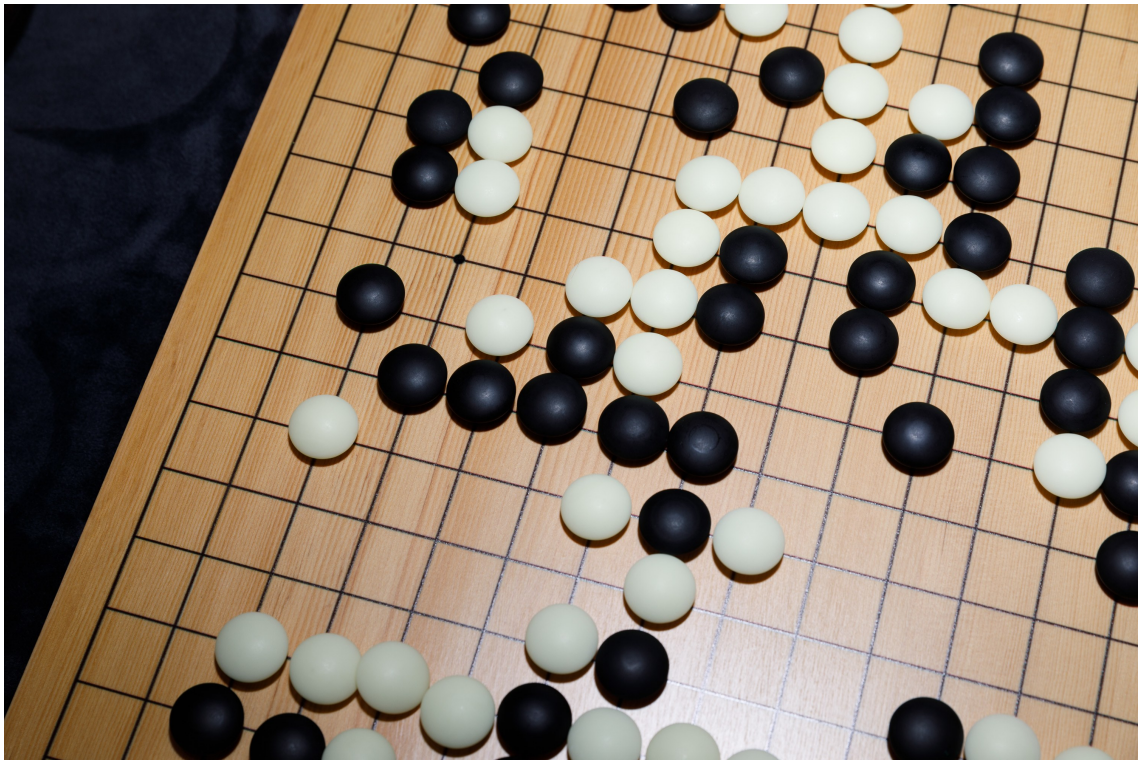
Google Maps


INOVA USP - Centro de Inovação da US
Escola de Artes, Ciências e Humanidad
Adicionar destino
Sair agora
OPÇÕES
Envie rotas para seu smartphone
via BR-116
41 min
Trajeto mais rápido agora devido às condições de trânsito
36,3 km
DETALHES
via Corredor Norte-Sul
56 min
40,6 km
22:23 - 00:17 (sexta-feira)
1 h 54 min
8012-10
METRÔ L4
METRÔ L3
CPTM L12
2735-10
Conheça Escola de Artes, Ciências e Humanidades da Universidade de São Paulo
Restaurantes
Hotéis
Postos de gasolina
Mais
Fazer login
41 min
36,3 km
56 min
40,6 km
INOVA USP - Centro de Inovação da USP
Escola de Artes, Ciências e...
São Paulo
Guarulhos
Osasco
Carapicuíba
Satélite
Google

Atari



Go



Starcraft

Processo Markoviano de Decisão

Valdinei Freire

(EACH - USP)

Definição

Um processo markoviano de decisão (*Markovian Decision Process* – MDP) é definido por uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$ onde:

- $s \in \mathcal{S}$ são estados possíveis;
- $a \in \mathcal{A}$ são ações possíveis;
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ é a função de transição; e
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \Re$ é a função recompensa.

Um MDP define o seguinte processo para todo tempo $t \geqslant 0$:

1. o agente percebe o estado s_t;

2. o agente escolhe e executa uma ação a_t;

3. o ambiente transita para o próximo estado s_{t+1} com probabilidade $\Pr(s_{t+1} = s' | a_t = a, s_t = s) = T(s, a, s')$; e

4. o agente recebe uma recompensa $R(s_t, a_t, s_{t+1})$.

Um agente que escolhe as ações a_t deve fazer isso de modo a buscar recompensas positivas e evitar recompensas negativas.

Estado Inicial

Pode-se também especificar um estado inicial:

- estado inicial único, simplesmente denominado s_0 (um MDP seria uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R, s_0 \in \mathcal{S} \rangle$)

- distribuição de probabilidade para o estado inicial, uma função de estado inicial $I : \mathcal{S} \rightarrow [0, 1]$, tal que $\Pr(s_0 = s) = I(s)$ (um MDP seria uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R, I \rangle$)

Horizontes

Quando o processo acaba?

- Horizonte finito: é definido um horizonte máximo N, de tal forma que o processo continua enquanto $t \leqslant N$.
- Horizonte infinito: o processo nunca para
- Horizonte indeterminado: considera-se estados absorvedores, usualmente um conjunto de estados metas $\mathcal{G}$, de tal forma que o processo acaba quando $s_t \in \mathcal{G}$

Estado Absorvedor

Um estado absorvedor é um estado que nunca sai dele e produz recompensa nula, isto é, para todo $s \in \mathcal{G}$ e $s' \neq s \in \mathcal{S}$, tem-se que $T(s, a, s) = 1$ e $T(s', a, s) = 0$ e $R(s, a, s) = R(s, a, s') = 0$.

Uma outra de forma é considerar que para todo $s \in \mathcal{G}$ e $s' \in \mathcal{S}$, tem-se que $T(s, a, s') = 0$.

Recompensas versus Custos

Quando as recompensas envolvidas são todas negativas e considera-se estados metas, tem-se o problema de Caminho Estocástico mais Curto (*Shortest Stochastic Path* – SSP). Nesse caso, a função recompensa é trocada por uma função custo $C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}^+$, nesse caso, um SSP é definido por $\langle \mathcal{S}, \mathcal{A}, T, C, s_0 \in \mathcal{S}, \mathcal{G} \rangle$.

Outras variações bastante encontradas nos trabalhos é com relação ao argumento da função recompensa. Nos trabalhos, é bastante comum encontrar as seguintes variações: $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ e $R : \mathcal{S} \rightarrow \mathbb{R}$.

Finalmente, pode-se considerar uma recompensa estocástica, na qual a recompensa r_t no tempo t vem de uma distribuição de probabilidade $\Pr(r_t = x | s_t = s, a_t = a, s_{t+1} = s')$.

Ações por Estado

Outra variação bastante comum em trabalhos de Planejamento ou Aprendizado por Reforço é restringir as ações disponíveis para cada estado, isto é, considera-se uma função $A : \mathcal{S} \rightarrow 2^{\mathcal{A}}$ e as ações escolhidas no tempo t são limitadas a $a_t \in A(s_t)$.

Observação Parcial

Uma formalização mais completa que o MDP e também mais realista para vários problemas é considerar observações parciais.

O agente observa o estado indiretamente por uma função de observação.

Dado um conjunto de observações $\mathcal{O}$, em cada tempo t o agente faz uma observação $o_t \in \mathcal{O}$. A relação da observação o_t com o estado s_t é dada por uma função de observação $O : \mathcal{S} \rightarrow (\mathcal{O} \rightarrow [0, 1])$ ou $O : \mathcal{S} \times \mathcal{O} \rightarrow [0, 1]$, tal que $\Pr(o_t = o|s_t) = O(s, o)$.

Políticas

As ações a_t podem ser escolhidas de forma arbitrária e são definidas por uma política de ações π, ou, simplesmente, política.

Essencialmente, um política mapeia situações em uma ação.

Políticas podem ser classificadas sob três aspectos:

- estacionária ou não-estacionária
- determinista ou probabilística
- markoviana ou não-markoviana

Políticas

Estacionária versus não-estacionária

- Uma política estacionária é igual para todo tempo $t \geqslant 0$, enquanto uma política não-estacionária vária com o tempo.

Determinista versus probabilística

- Uma política determinista escolhe sempre a mesma ação quando em uma mesma situação, enquanto uma política probabilística escolhe uma ação segundo uma distribuição de probabilidade.

Políticas

A execução de uma política associada ao processo regido por uma função de transição T, gera histórias $h_t = s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t$.

Markoviana versus Não-markoviana

- Uma política markoviana depende apenas do estado atual, isto é, escolhe a ação a_t com base apenas no estado atual s_t, enquanto uma política não-markoviana escolhe a ação a_t com base no histórico h_t.

Políticas

Política estacionária, markoviana e determinista:
$\pi : \mathcal{S} \to \mathcal{A}$.

Política não-estacionária, markoviana e determinista:
$\pi : \mathcal{S} \times \mathbb{N} \to \mathcal{A}$.

Política estacionária, não-markoviana e determinista:
$\pi : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^* \times \mathcal{S} \to \mathcal{A}$.

Política estacionária, markoviana e probabilística:
$\pi : \mathcal{S} \to (\mathcal{A} \to [0, 1])$ ou $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

Avaliando Políticas

Ao executar uma política π um agente escolhe ações $a_t = \pi(s_t)$ e gera um histórico $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$, que pode ser finita ou infinita, dependendo do tipo de horizonte considerado. Há três formas básicas na literatura para avaliar uma política: média, somatória e desconto.

Se considerarmos todos os estamos como potenciais estados iniciais, a avaliação de uma política consiste em criar uma função valor $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$ que especifica um escalar para cada estado. Dessa forma, dado um estado inicial s_0 arbitrário, pode-se comparar duas políticas π e π' comparando os valores $V^\pi(s_0)$ e $V^{\pi'}(s_0)$.

Recompensa Média

Definition 1. O valor de uma política segundo o critério de recompensa média é dado por:

$$V^{\pi}(s) = \lim_{N \to \infty} \frac{1}{N} \mathsf{E}\left[\sum_{t=0}^{N-1} r_t \middle| s_0 = s, \pi \right].$$

A recompensa média pode ser utilizada apenas para horizontes infinitos, no qual o processo nunca acaba. Note que o horizonte infinito é especificado pelo limite $N \to \infty$. Ainda, note que se as recompensas se extinguirem, como no caso de horizonte indeterminado, o limite tenderá a zero para qualquer política.

Recompensa Acumulada Esperada

Definition 2. O valor de uma política segundo o critério de recompensa acumulada esperada é dado por:

$$V^{\pi}(s) = \mathsf{E}\left[\sum_{t=0}^{\infty} r_t \middle| s_0 = s, \pi\right].$$

A recompensa acumulada esperada pode ser utilizada apenas para horizontes finitos ou indeterminados. Como todas as recompensas são acumuladas, esse valor será finito apenas se as recompensas cessarem. Dessa forma, esse critério não pode ser utilizado para horizontes infinitos.

Recompensa Acumulada Descontada

Definition 3. O valor de uma política segundo o critério de recompensa acumulada descontada é dado por:

$$V^{\pi}(s) = \mathsf{E}\left[\sum_{t=0}^{\infty} \gamma^{t} r_t \middle| s_0 = s, \pi\right],$$

onde $\gamma < 1$ é um fator de desconto.

A recompensa acumulada descontada garante que, quando a recompensa é limitada, o acumulo descontado seja sempre finito, independente do tipo de horizonte.

Além disso, quando $\gamma \to 1$, esse critério se aproxima do critério de recompensa média se o horizonte for infinito ou se aproxima do critério de recompensa acumulada se o horizonte for finito ou indeterminado.