Aprendizagem por Reforço II

slides baseados na disciplina CS188 de UC Berkeley

OpenAI

About Blog

About OpenAI

OpenAI is a non-profit artificial intelligence research company. Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return.

In the short term, we're building on recent advances in AI research and working towards the next set of breakthroughs.

Aprendizagem por Reforço

Processos de Decisão de Markov (MDP).

- Conjunto de estados $s \in S$
- Conjunto de ações $a \in A$
- Função de transição $T(s, a, s')$ ou $P(s'|s, a)$
- Função de recompensa $R(s, a, s')$ (e fator γ)
- Um estado inicial
- Objetivo ainda é encontrar política π^*

Diferença: $T(s, a, s')$ e $R(s, a, s')$ são desconhecidas.

Mapa até o Momento

MDP Conhecido: Solução Offline

Objetivo	Técnica
Calcular V*, Q*, π*	Iteração de Valor / política (IV/IP)
Avaliar uma política fixa π	Avaliação de Política (AP)

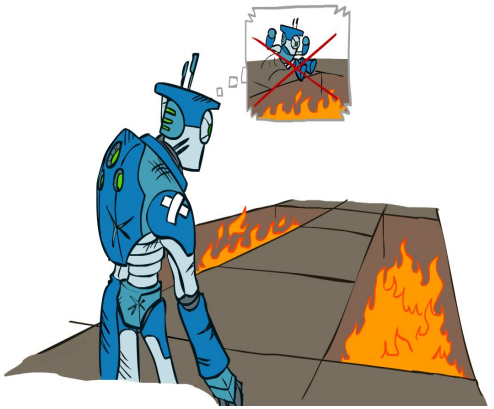
MDP Desconhecido: Baseado em Modelo

Objetivo	Técnica
Calcular V*, Q*, π*	IV/IP em MDP aprox.
Avaliar uma política fixa π	AP em MDP aprox.

MDP Desconhecido: Livre de Modelo

Objetivo	Técnica
Calcular V*, Q*, π*	Aprendizagem-Q
Avaliar política fixa π	TD

Arrependimento

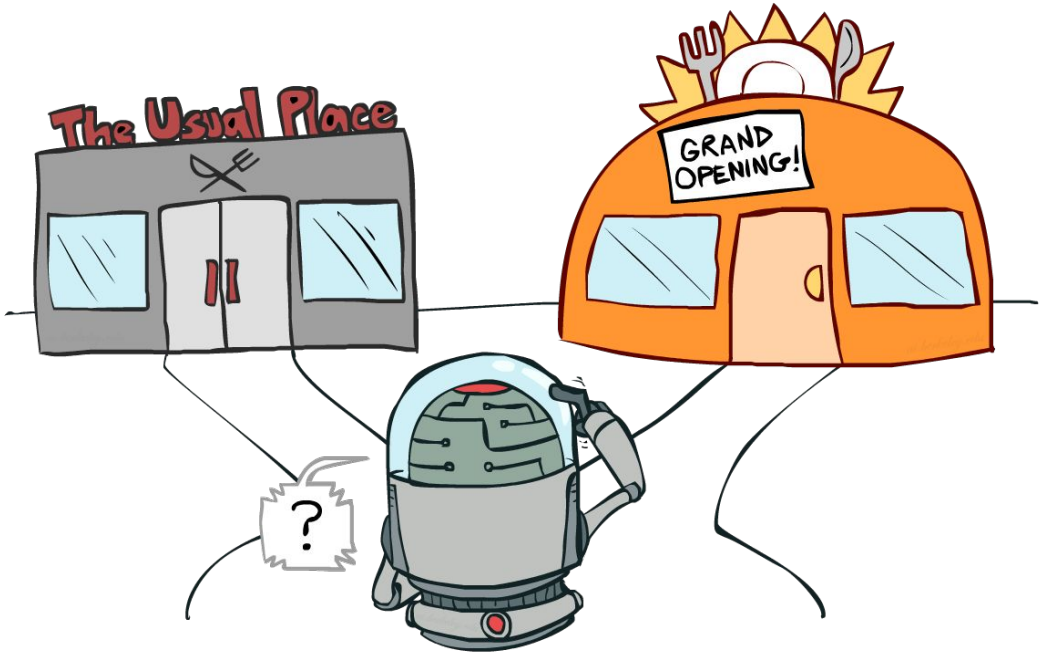


Mesmo executando a política ótima o agente pode cometer erros.

Arrependimento de π mede a diferença de recompensas obtidas com π^* com as recompensas obtidas com π.

Minimizar arrependimento é mais difícil que aprender π^*, pois requer aprender π^* de forma ótima.

Exploração vs. Exploitação



Como Explorar?

Método simples: ϵ-guloso.

- Com uma probabilidade pequena ϵ, escolha uma ação aleatória.
- Escolha de forma gulosa com probabilidade 1-ϵ

Vantagem
Eventualmente explora todo o espaço (requerimento de Aprendizagem-Q)

Desvantagem
Toma decisões sub-ótimas mesmo depois de ter aprendido.

Soluções
Reduzir o valor de ϵ ao longo do tempo.
Utilizar funções de exploração

Já vimos isso antes.

Função de Exploração

- Explorar apenas aqueles estados que são mais desconhecidos.
- Utilizar função que considera o valor estimado u, o número de vezes que o par (s, a) foi visitado n, e uma constante k:

$$f(u, n) = u + \frac{k}{n}$$

Amostra regular Aprendizagem-Q: $$amostra = R(s, a, s') + \gamma \max_{a'} Q(s', a')$$

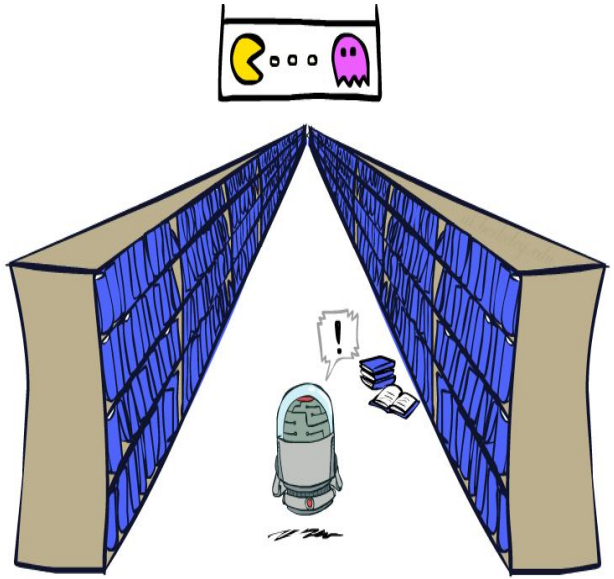
Amostra modificada Aprendizagem-Q:

$$amostra = R(s, a, s') + \gamma \max_{a'} f(Q(s', a'), N(s', a'))$$

O Q-learning funciona bem quando temos poucos estados.

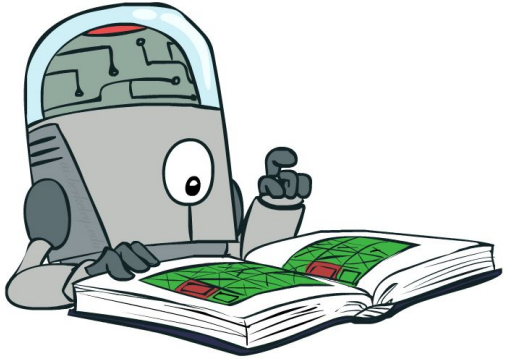
E se o número de estados for grande?

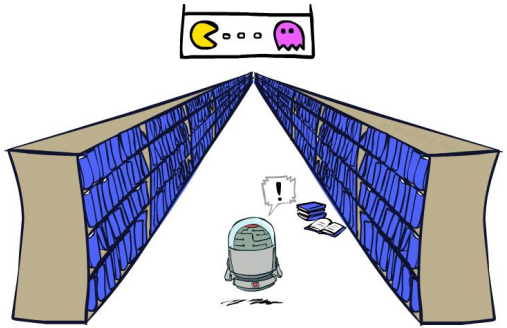
Mundo real: carro autônomo.



Generalização

- Aprendizagem-Q utiliza uma tabela com cada estado do espaço de busca.
- Não realístico para problemas interessantes.
  - Impossível visitar todos estados
  - Impossível armazenar todos estados na memória





Generalizar: aprender com poucos estados e generalizar para estados não vistos.

Generalização

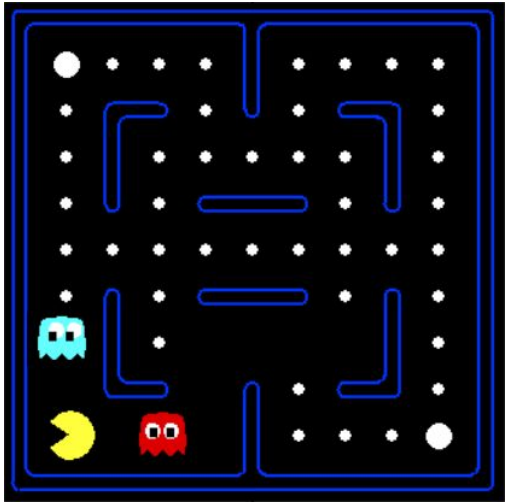
Através de experiência descobrimos que esse estado é ruim

Em Aprendizagem-Q com uma tabela não sabemos

Ou nem mesmo esse!

Representação com Atributos

- Utilizar um vetor de atributos (propriedades) para descrever os estados.



Exemplos:

Distância para o fantasma mais próximo.
Distância para a pílula mais próxima.
Número de fantasmas na vizinhança.
Número de pílulas restantes.
etc…

Atributos podem também descrever um estado-q.
(ex.: ação que nos leva para mais próximo de pílulas)

Representação Linear

- Utilizando a representação de atributos, podemos reescrever V(s) e Q(s, a) da seguinte forma.

$$V(s) = w_1 f_1(s) + w_2 f_2(s) + \ldots + w_n f_n(s)$$

$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \ldots + w_n f_n(s,a)$$

- **Vantagem**: aprendizagem acontece de forma mais rápida, pois a representação é mais compacta.
- **Desvantagem**: pode haver estados com atributos parecidos mas com valores V e Q diferentes.

Aprendizagem-Q Aproximado

$$Q(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \ldots + w_n f_n(s,a)$$

Em **Aprendizagem-Q Tabular** com uma tupla (s, a, r, s') calculamos:

$$diferenca = \left[r + \gamma \max_{a'} Q(s',a')\right] - Q(s,a)$$

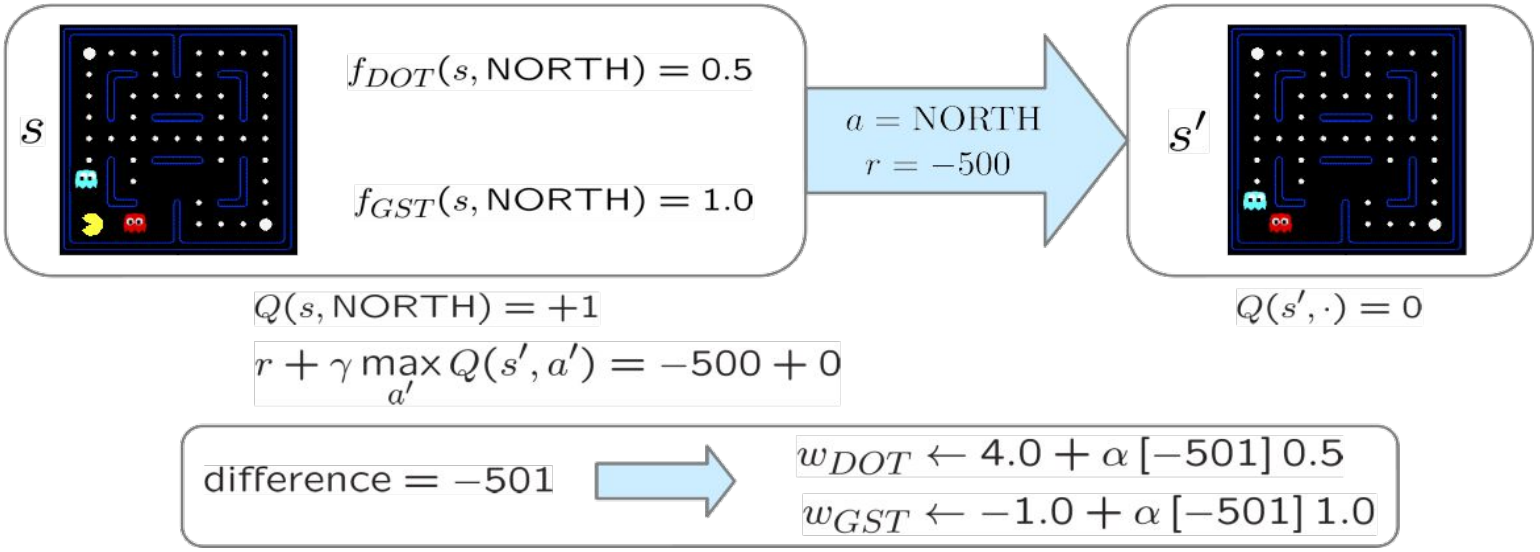
$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[diferenca\right]$$

Em **Aprendizagem-Q Aproximado**, fazemos:

$$w_i \leftarrow w_i + \alpha\left[diferenca\right] f_i(s,a)$$

Exemplo Aprendizagem-Q Aproximada

$$Q(s,a) = 4.0 f_{DOT}(s,a) - 1.0 f_{GST}(s,a)$$



Para α=0.01 temos w_{DOT}=4.0+0.01×-501×0.5=1.495 e w_{GST}=-1.0+0.01×-501×1=-6.01

$$Q(s,a) = 1.5 f_{DOT}(s,a) - 6 f_{GST}(s,a)$$

FIM