

<div> <div> <div>Inteligência Artificial</div> <div> <div>Luís A. Alexandre</div> <div>UBI</div> <div>Ano lectivo 2019-20</div> </div> </div> <div> <div>1 / 34</div> </div> </div>	
Luís A. Alexandre (UBI)	Inteligência Artificial

<div> <div> <div>Conteúdo</div> </div> </div>	
<div> <div>Aprendizagem por reforço</div> <div>Introdução</div> <div>Elementos da Aprendizagem por Reforço</div> <div>Bandido com n-braços</div> </div>	<div> <div>Exemplos da Aprendizagem por Reforço</div> <div>Aplicações da Aprendizagem por Reforço</div> <div>Leitura recomendada</div> </div>
Luís A. Alexandre (UBI)	Inteligência Artificial

<div> <div> <div>Introdução</div> </div> </div>	
<div> <div> <div> <div>► Nesta aula vamos estudar a Aprendizagem por Reforço (AR) onde o agente terá que aprender muito à semelhança do que acontece com os humanos: com o resultado das suas ações.</div> <div>► Imaginemos que o nosso agente tem que aprender a jogar xadrez, mas sem receber feedback de um professor.</div> <div>► O principal problema reside em o agente perceber que um checkmate é uma coisa boa quando é ele que o faz e má quando o oponente lhe faz a ele.</div> <div>► O ponto essencial da AR é receber este feedback, chamado o reforço, para o agente poder guiar a sua aprendizagem.</div> </div> </div> </div>	
Luís A. Alexandre (UBI)	Inteligência Artificial

<div> <div> <div>Introdução</div> </div> </div>	
<div> <div> <div> <div>► Note-se que este reforço, no caso do xadrez, só chega no final do jogo. Noutros problemas o reforço pode ser recebido com maior frequência.</div> <div>► No ping-pong cada bola jogada fornece informação de reforço ao agente pois sabe logo se ganhou ou não um ponto.</div> <div>► Quando temos ambientes complexos, a AR é por vezes a única forma de treinar um agente.</div> </div> </div> </div>	
Luís A. Alexandre (UBI)	Inteligência Artificial

<div> <div> <div>Introdução</div> </div> </div>	
<div> <div> <div> <div>► Para ter um desempenho bom o agente não pode preocupar-se apenas com o resultado imediato das suas ações: é preciso levar em conta as consequências a longo prazo das ações.</div> <div>► Exemplo: para maximizar o nosso ordenado futuro é melhor ir agora para a escola embora a recompensa monetária de frequentar a escola seja negativa (pagar propinas).</div> <div>► Assim a AR é particularmente adaptada para problemas em que exista um equilíbrio a alcançar entre recompensas a curto e a longo prazo.</div> </div> </div> </div>	
Luís A. Alexandre (UBI)	Inteligência Artificial

<div> <div> <div>Conteúdo</div> </div> </div>	
<div> <div>Aprendizagem por reforço</div> <div>Introdução</div> <div>Elementos da Aprendizagem por Reforço</div> <div>Bandido com n-braços</div> </div>	<div> <div>Exemplos da Aprendizagem por Reforço</div> <div>Aplicações da Aprendizagem por Reforço</div> <div>Leitura recomendada</div> </div>
Luís A. Alexandre (UBI)	Inteligência Artificial

Introdução

- ▶ Na AR o agente vai interagir com o ambiente em passos de tempo discretos.
- ▶ Em cada passo t o agente recebe uma **observação** do ambiente b que inclui a **recompensa** r relativa à ação que realizou no instante anterior $(t - 1)$.
- ▶ De seguida, deve escolher uma nova **ação** a realizar, $a(t)$, de entre o conjunto de ações possíveis. Esta ação é enviada para o ambiente.
- ▶ O ambiente passa então a um **novo estado** $s(t + 1)$ e é definida a recompensa relativa à ação $a(t)$, que será $r(t + 1)$.
- ▶ Existe então a seguinte **transição** de estado do ambiente, provocada pela ação do agente: $(s(t), a(t), s(t + 1))$.
- ▶ O **objetivo** do agente é recolher o máximo de recompensa possível.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

7 / 34

Política

- ▶ Nalguns casos a política pode ser uma simples tabela: se estou no estado x devo fazer a ação y .
- ▶ Noutros casos podemos ter que realizar bastantes cálculos, incluindo até fazer pesquisa para decidirmos que ação executar.
- ▶ No caso mais geral a política pode ser **estocástica**: num dado estado a ação a executar depende duma distribuição probabilística.
- ▶ A política poderá ser aprendida com o desenrolar da ação (ver slide seguinte).

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

9 / 34

Função Valor

- ▶ Ao contrário da função recompensa que nos indica o que é bom em termos imediatos, a função valor indica o que será bom no longo prazo.
- ▶ O **valor de um estado** leva em conta os estados que provavelmente se lhe seguirão e as recompensas que se podem obter também nesses estados.
- ▶ Exemplo: estudar hoje pode parecer que tem fraca recompensa, mas tem valor, pois no longo prazo trará melhores possibilidades de obter bons empregos.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

11 / 34

Elementos da Aprendizagem por Reforço

- ▶ Além do **agente** e do **ambiente** onde ele se encontra, existem 4 elementos fundamentais na AR:
 - ▶ **política**: define o comportamento do agente. Indica que ação tomar quando o agente se encontra num dado estado.
 - ▶ **função recompensa** de um estado: indica qual a recompensa que está associada a esse estado.
 - ▶ **função valor** de um estado: indica qual o valor total de recompensa esperado se o agente partir deste estado.
 - ▶ **modelo do ambiente**: este é o único elemento que não é obrigatório, mas existe em muitos casos.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

8 / 34

Função Recompensa

- ▶ É a FR que define qual é o **objetivo** do problema de AR.
- ▶ Mapeia um par (estado,ação) para um valor, a recompensa, que indica quão desejável é o estado que resulta da aplicação da ação ao estado.
- ▶ O objetivo do agente é apenas o de maximizar a recompensa total que obtém **no longo prazo**.
- ▶ A FR define o que são boas e más ações.
- ▶ A FR para um animal poderia ser o prazer e a dor: são os resultados imediatos das ações do animal.
- ▶ A FR tem que ser fixa mas pode ser usada para **alterar a política**: se uma dada ação que a política mandou executar num dado estado recebe uma recompensa baixa, pode ajustar-se a política para que seja usada outra ação nesse estado.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

10 / 34

Modelo do ambiente

- ▶ O modelo poderá ser usado para, partindo de um dado estado e duma possível ação, tentar **prever** o estado resultante e a respetiva recompensa.
- ▶ Os modelos são usados para **planeamento**: decidir o conjunto de ações a tomar tendo em vista estados futuros, antes de eles serem efetivamente vivenciados.
- ▶ A AR mais simples não usa modelos, apenas tentativa e erro.
- ▶ As abordagens mais avançadas podem usar ambos: aprender por tentativa e erro, criar um modelo e passar a usá-lo para planear as ações futuras.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

12 / 34

Tipos de feedback

- ▶ Podemos considerar que existem 2 tipos de feedback que um sistema pode receber:
 - ▶ **avaliativo**: indica se a ação executada foi ou não boa (mas não diz se é a melhor ou a pior possível)
 - ▶ **instrutivo**: indica qual é a ação correta a executar (é aprendizagem supervisionada)
- ▶ Na AR só temos feedback avaliativo.
- ▶ Vejamos um problema simplificado de AR: o bandido de n -braços.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

13 / 34

Bandido com n -braços

- ▶ Os ingleses chamam a uma slot machine um bandido com 1-braço.
- ▶ Nós vamos estudar o problema se a slot machine tivesse n braços em vez de apenas um.
- ▶ Descrição formal do problema:
 - ▶ O agente deve **escolher** uma entre n ações possíveis.
 - ▶ Após cada ação recebe uma **recompensa** numérica que tem valor obtido duma distribuição de probabilidade estacionária (não varia com o tempo).
 - ▶ O objetivo é **maximizar** o total de recompensas num dado período, por exemplo, 1000 ações.
 - ▶ Chamamos uma **jogada** a cada escolha de uma ação.



Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

15 / 34

Exploration-exploitation

- ▶ Ao escolher a ação gulosa estamos a usufruir ao máximo do que temos (**exploitation**).
- ▶ Ao escolher outra ação que não a gulosa estamos a explorar (**exploration**) pois estamos a procurar estimativas para as outras ações que eventualmente poderão levar-nos a encontrar uma melhor que a gulosa atual.
- ▶ Se só tivéssemos uma jogada para realizar antes do fim do jogo, o melhor seria seguir a abordagem da exploitation.
- ▶ Mas com muitas jogadas é importante fazer exploration pois permite procurarmos melhores jogadas que a que atualmente se apresenta como a melhor.
- ▶ Como numa dada jogada temos que escolher entre exploration e exploitation, existe um **conflito** entre as duas abordagens.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

17 / 34

Conteúdo

Aprendizagem por reforço

Introdução
Elementos da Aprendizagem por Reforço

Exemplos da Aprendizagem por Reforço
Aplicações da Aprendizagem por Reforço
Leitura recomendada

Bandido com n -braços

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

14 / 34

Bandido com n -braços

- ▶ A ideia será descobrirmos quais os braços que dão maior recompensa e focarmos as nossas jogadas nesses braços.
- ▶ Cada jogada tem uma recompensa média associada: o valor em média que se recebe se escolhermos acionar aquele braço.
- ▶ Vamos chamar a este valor médio o **valor da ação**.
- ▶ Se soubéssemos os valores associados a todas as ações, o problema estaria resolvido: só acionávamos o braço com maior valor.
- ▶ Vamos criar **estimativas** dos valores associados a cada ação.
- ▶ Em qualquer momento temos uma ação que tem o maior valor estimado: chamamos-lhe a **ação gulosa** (greedy).

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

16 / 34

Métodos ação-valor

- ▶ Vejamos formas de fazer a escolha entre exploration e exploitation.
- ▶ Temos que tentar estimar os valores das ações.
- ▶ Vamos chamar ao **verdadeiro valor da ação** a , $Q^*(a)$.
- ▶ Chamamos à nossa **estimativa do valor da ação** a , ao fim de t jogadas, $Q_t(a)$.
- ▶ Como o valor da ação a é a recompensa média que se obtém quando se escolhe essa ação, uma forma simples de estimar esse valor é o cálculo da média das recompensas obtidas quando se escolheu essa ação:

$$Q_t(a) = \frac{1}{k_a} \sum_{i=1}^{k_a} r_i \quad (1)$$

- onde k_a é o número de vezes que se escolheu a ação a e r_i é a recompensa recebida em cada em cada escolha.
- ▶ Se $k_a = 0$ definimos que $Q_t(a) = 0$.
 - ▶ Conforme aumentamos k_a a estimativa $Q_t(a)$ vai tender para o verdadeiro valor $Q^*(a)$.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

18 / 34

Métodos ação-valor

- ▶ Como usar a nossa estimativa para escolher as ações?
- ▶ A forma mais básica seria escolher a que tiver a maior estimativa de valor.
- ▶ O problema é que assim não fazemos exploration, apenas exploitation.
- ▶ Para resolver isso podemos dizer que numa dada proporção das jogadas ϵ , em vez de escolhermos a opção gulosa, escolhemos uma outra ao acaso.
- ▶ Esta é chamada a abordagem ϵ -gulosa.
- ▶ Com esta abordagem, garantimos que com o aumento do número de jogadas, iremos ficar a conhecer excelentes estimativas dos verdadeiros valores de $Q^*(a)$ para todas as ações a .

Luís A. Alexandre (UBI)

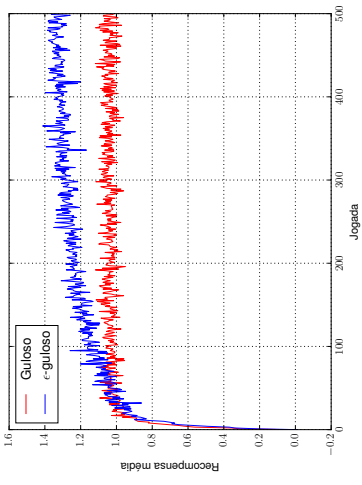
Inteligência Artificial

Ano lectivo 2019-20

19 / 34

Exemplo

- ▶ Recompensa média para 2 agentes, um guloso e um $\epsilon = 0.1$ guloso. 10 braços.
- ▶ Problema com 500 jogadas, valor médio de 1000 repetições de cada experiência.



Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

20 / 34

Bandido com n-braços

- ▶ O problema do bandido com n -braços é uma versão simplificada do problema geral da AR.
- ▶ A simplificação ocorre a três níveis:
 - ▶ As distribuições de probabilidades das recompensas de cada ação são estacionárias;
 - ▶ Não precisamos de aprender uma política: as ações a executar não dependem do estado em que nos encontramos;
 - ▶ Cada ação só afeta a recompensa imediata e não as recompensas futuras.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

21 / 34

Bandido com n-braços não estacionário

- ▶ No caso **não estacionário** devemos alterar a forma como fazemos a estimativa das recompensas.
- ▶ Uma forma de o fazer será pesar mais as recompensas mais recentes que as mais antigas.
- ▶ Deste modo, podemos alterar a eq. (1) e passar a usar a seguinte: (2)

$$Q_t = Q_{t-1} + \alpha(r_t - Q_{t-1})$$
 onde $\alpha \in (0, 1]$ é uma constante chamada o **passo**.
- ▶ Isto na prática torna a nossa estimativa uma média pesada das recompensas em vez de ser uma média simples.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

22 / 34

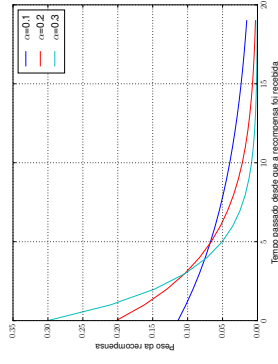
Bandido com n-braços não estacionário

- ▶ Vejamos:

$$\begin{aligned}
 Q_t &= Q_{t-1} + \alpha(r_t - Q_{t-1}) \\
 &= \alpha r_t + (1 - \alpha) Q_{t-1} \\
 &= \alpha r_t + (1 - \alpha)(Q_{t-2} + \alpha(r_{t-1} - Q_{t-2})) \\
 &= \alpha r_t + (1 - \alpha)Q_{t-2} + (1 - \alpha)\alpha r_{t-1} - (1 - \alpha)\alpha Q_{t-2} \\
 &= \alpha r_t + (1 - \alpha)^2 Q_{t-2} + (1 - \alpha)\alpha r_{t-1} \\
 &= \dots \\
 &= Q_0(1 - \alpha)^t + \alpha \sum_{i=0}^{t-1} (1 - \alpha)^i r_{t-i}
 \end{aligned} \tag{3}$$

Bandido com n-braços não estacionário

- ▶ Daqui podemos concluir que temos uma soma das recompensas, pesada pelo termo $\alpha(1 - \alpha)^{t-i}$ para a recompensa r_i .
- ▶ Quanto mais tempo tiver passado desde que recebemos a recompensa r_i menos será o peso que ela tem na nossa estimativa da recompensa média:



Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

23 / 34

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

24 / 34

Bandido com n -braços não estacionário com política

- ▶ Para modificarmos o problema do bandido para termos que **aprender uma política**, consideremos que a slot machine pode mudar as suas recompensas de cada vez que é feita uma ação, indicando o modo em que se encontra através da cor do ecrã.
- ▶ Agora o nosso agente tem que aprender a estimar as recompensas para cada braço da máquina mas essas recompensas variam com a cor do ecrã.
- ▶ Por exemplo: se a cor for vermelho o braço que em média está a dar maiores recompensas pode ser o 3, mas se a cor for azul o melhor braço pode ser o 1.
- ▶ Isto obriga o nosso agente a aprender uma política: qual a melhor ação dependendo do estado em que a máquina está.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

25 / 34

Conteúdo

Aprendizagem por reforço

Introdução

Elementos da Aprendizagem por

Reforço

Bandido com n -braços

Exemplos da Aprendizagem por

Reforço

Aplicações da Aprendizagem por

Reforço

Leitura recomendada

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

27 / 34

Robô para pick-and-place

- ▶ A tarefa pick-and-place consiste no pegar num objeto que se encontra num local pré-definido e colocá-lo noutra local também pré-definido.
- ▶ O nosso agente seria o robô e o problema de AR aparece quando pretendemos obter movimentos no braço que sejam simultaneamente simples e rápidos.
- ▶ O agente tem que controlar os motores que fazem mexer os vários componentes do braço e recebe informação relativa à posição e velocidade destes componentes no espaço 3D (esta informação define o **estado**).
- ▶ A **recompensa** poderá ser +1 se o objeto for movido corretamente e, para encorajar movimentos suaves, podemos atribuir uma recompensa negativa pequena a movimentos bruscos.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

29 / 34

Bandido com n -braços não estacionário com política

- ▶ Como fazer para aprender a política neste caso?
- ▶ A probabilidade de, no instante t , estar no estado s e escolher a ação a , é a **política** $\pi_t(s, a)$.
- ▶ A política vai sendo ajustada de acordo com a experiência do agente.
- ▶ Para o exemplo do bandido, poderíamos aprender a política simplesmente aprendendo as estimativas de recompensas, em função da cor do ecrã: teríamos uma distribuição de estimativas para o ecrã vermelho e outra para o azul (que são os estados).
- ▶ Uma possível política seria usar a ação com maior estimativa $1-\epsilon$ por cento das vezes, dependendo do estado. Nos restantes casos, usar uma ação aleatória.
- ▶ Este contexto está já muito próximo do problema geral da AR: só falta que a ação que escolhemos num momento pudesse afetar não só a recompensa imediata mas também as recompensas da próxima ação para estarmos no caso mais geral.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

26 / 34

Bio-reator

- ▶ Um bio-reator é um contentor com nutrientes e bactérias usado na produção de produtos químicos úteis.
- ▶ Neste exemplo o nosso agente é responsável por definir a temperatura e a agitação do bio-reator.
- ▶ A temperatura é alterada com recurso a aquecedores e a agitação com recurso a motores.
- ▶ O **estado** do sistema é obtido com informação proveniente de termómetros e de outros sensores que fornecem informação relativa à quantidade de nutrientes e de outros químicos presentes no bio-reator.
- ▶ A **recompensa** poderá ser uma medida da taxa a que os químicos estão a ser produzidos.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

28 / 34

Robô para reciclagem

- ▶ Neste exemplo consideramos um robô que percorre um escritório e recolhe latas para reciclagem.
- ▶ Tem sensores para detetar as latas e um braço para lhes pegar e guardar num contentor que traz consigo.
- ▶ A AR surge quando o robot tem que decidir entre 3 possíveis **ações**:
 - ▶ procurar por latas
 - ▶ esperar que lhe tragam latas
 - ▶ ir recarregar as suas baterias
- ▶ O seu **estado** é dado pelo nível de carga da bateria.
- ▶ A **recompensa** pode ser: zero normalmente; positiva quando recolhe uma lata; negativa quando o nível de bateria baixa muito.

Luís A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

30 / 34

Aprendizagem por reforço

Aplicações da Aprendizagem por Reforço

Conteúdo

Aprendizagem por reforço

Introdução

Elementos da Aprendizagem por Reforço

Bandido com n -braços

Exemplos da Aprendizagem por Reforço

Aplicações da Aprendizagem por Reforço

Leitura recomendada

Luis A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

31 / 34

Aprendizagem por reforço

Aplicações da Aprendizagem por Reforço

Aplicações da Aprendizagem por Reforço

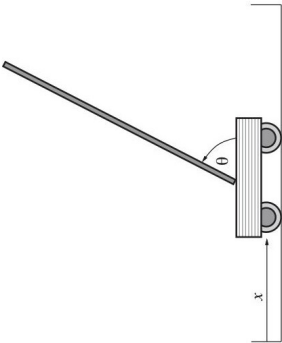


Imagem de Russell e Norvig

▶ Pêndulo invertido ou cart-pole.

▶ Exemplo clássico: o agente só mexe o carro para a esquerda ou direita e tem que manter o eixo na vertical.

▶ A informação recebida: $x, \theta, \frac{dx}{dt}, \frac{d\theta}{dt}$.

Luis A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

32 / 34

Aprendizagem por reforço

Aplicações da Aprendizagem por Reforço

Aplicações da Aprendizagem por Reforço

▶ Manobra nose-in circle (muito difícil).

▶ Controlo remoto do helicóptero com resultados muito superiores aos dos humanos a usarem o controlo remoto.

▶ Foto sobreposta das várias posições do helicóptero.



Foto de Russell e Norvig

Luis A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

33 / 34

Aprendizagem por reforço

Aplicações da Aprendizagem por Reforço

Leitura recomendada

▶ **Reinforcement Learning: An Introduction**, Richard S. Sutton and Andrew G. Barto, MIT Press, 1998, capítulos 1, 2 e sec. 3.1.

▶ Russell e Norvig, sec. 21.1 e 21.6.

Luis A. Alexandre (UBI)

Inteligência Artificial

Ano lectivo 2019-20

34 / 34