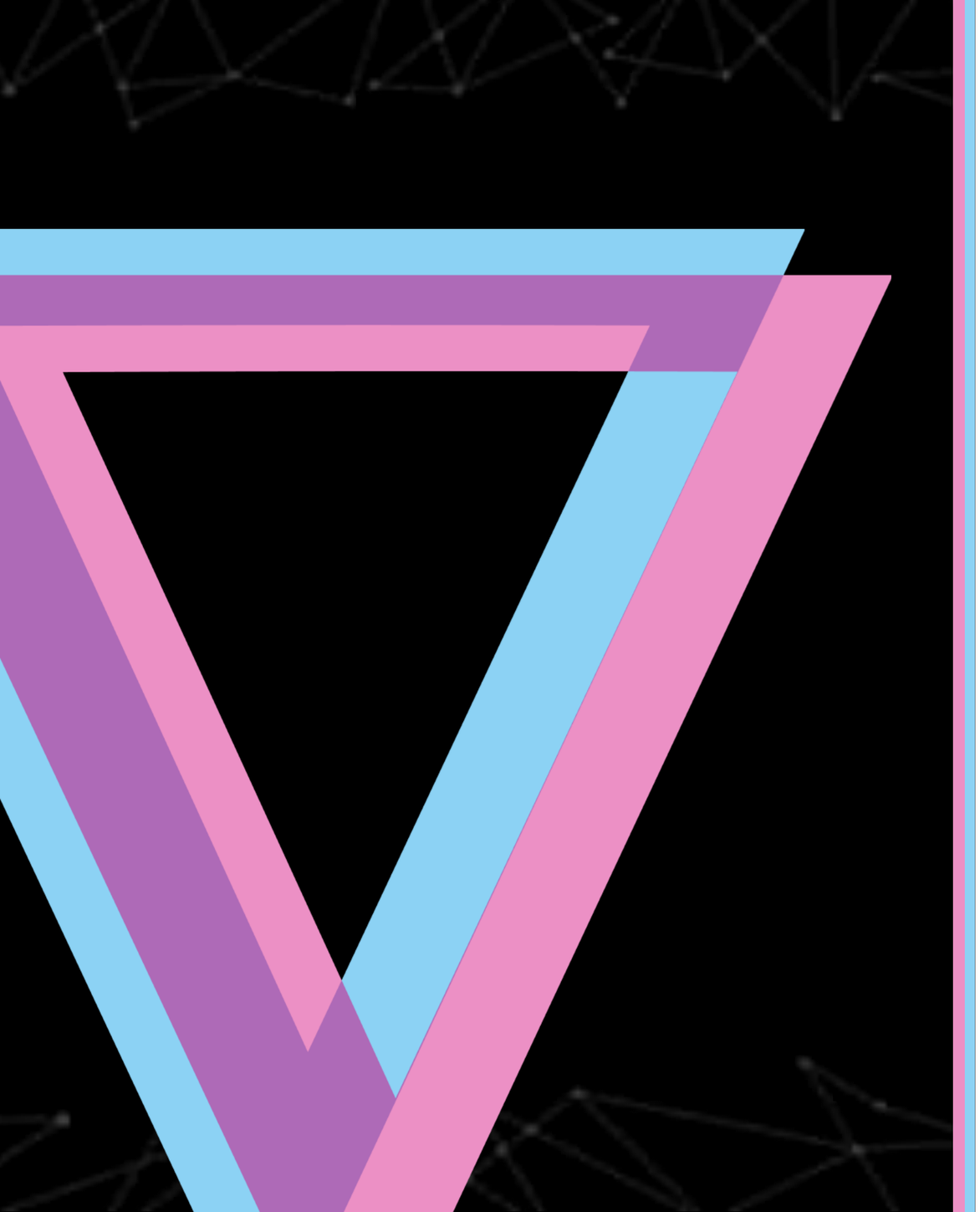


Reinforcement Learning

Glauco Fleury
@s/linkedin

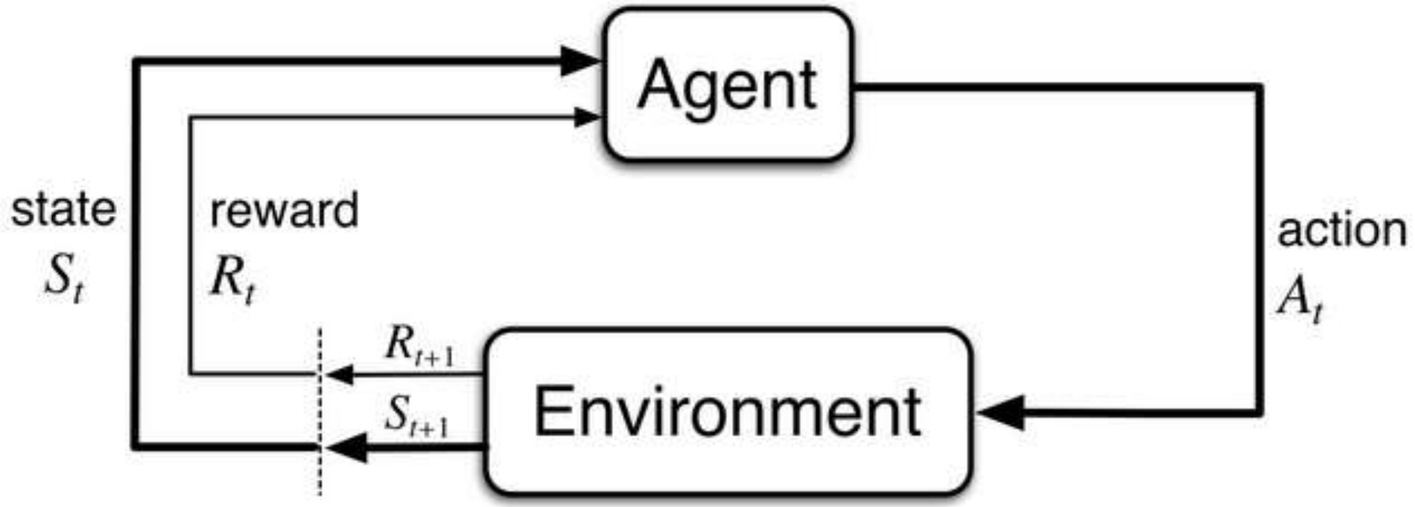
Presença

- Linktree: Presente na bio do nosso instagram
- Presença ficará disponível até 1 hora antes da próxima aula
- É necessário 70% de presença para obter o certificado

Presença

Recapitulando

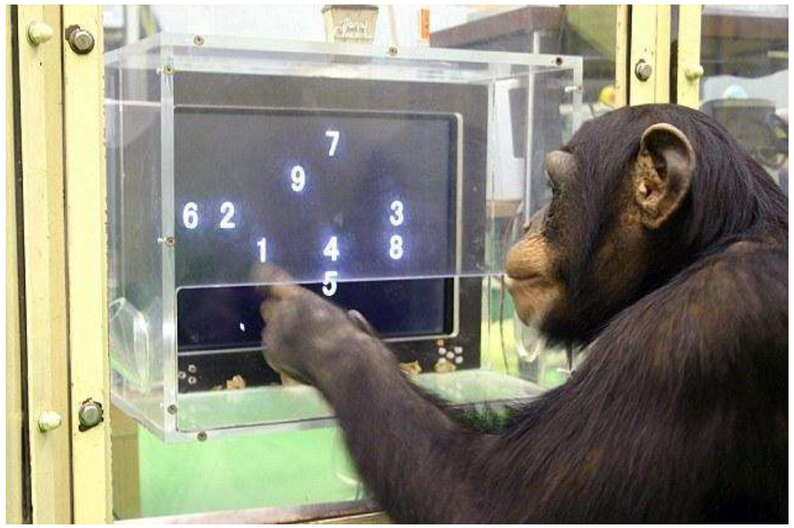
- Diagrama básico do ciclo Agente-Ambiente:



- $t \in \{1, 2, 3, ..\}$
- $s \in S$
- $a \in A(s)$
- $r \in \mathbb{R}, r \neq \infty$

Recapitulando

- Inspiração animal
- o modelo deve aprender enquanto experiência o mundo
- distinto de supervisionado e genético



Elementos / Ideia Geral

Elementos do RL

Quatro principais elementos:

- Policy
- Recompensa
- Função valor
- Modelo

Policy $\pi(a|s_t)$

- Estratégia do agente para lidar com o mundo
- Mapeamento entre o que o agente compreende do ambiente e suas ações
- Pode ser estocástica ou determinística

Retorno R_t

- Ganho ou perda total de uma interação
- Feedback imediato
- Ideal: caminho de maior recompensa a longo prazo, mesmo que imediatamente repreensivo
- Analogia: prazer e dor
- A recompensa **é o objetivo** do modelo

$$R_t = \sum_{k=0}^{T} (\gamma^k r_{t+k+1})$$

Função valor $V^{\pi}(s_t)$

- Cumulativa das recompensas de várias interações (longo prazo)
- Como o agente enxerga seu futuro em cada estado, seguindo uma política específica
- Deve ser estimado para problemas práticos (determiná-lo exatamente é custoso/impossível)

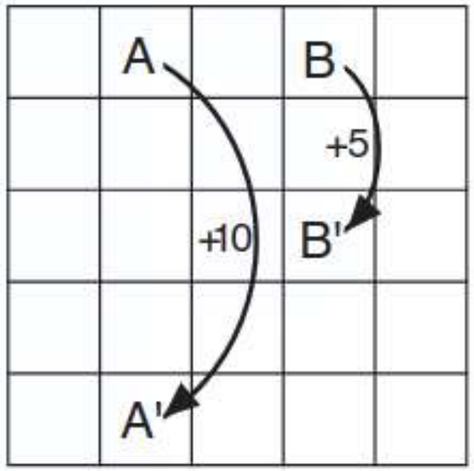
$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_t | S_t = s]$$

Modelo do ambiente

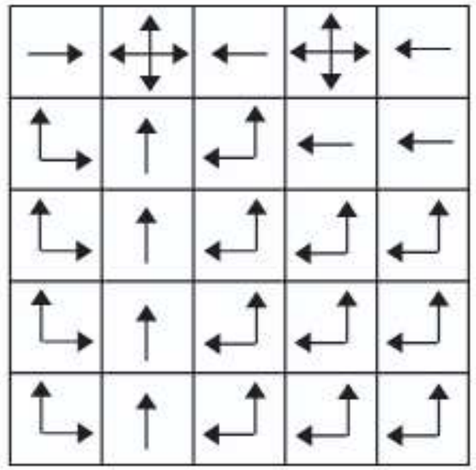
- Opcional
- Capacidade do agente de simular sua trajetória adiante, dadas suas escolhas agora (-tentativa e erro)
- Tentativa de predizer a melhor função valor, assim acelerando o aprendizado
- ex: rede neural acoplada a um agente

Exemplo: GridWorld

Objetivo: otimizar Policy e Função Valor
(como? Descubra na próxima aula)



22.0	24.4	22.0	19.4	17.5
19.8	22.0	19.8	17.8	16.0
17.8	19.8	17.8	16.0	14.4
16.0	17.8	16.0	14.4	13.0
14.4	16.0	14.4	13.0	11.7



N-Armed-Bandit

Contextualização

- Algoritmos iniciais e mais básicos, que motivaram o Reinforcement Learning atual
- “Bandit” = slot machine / máquina de cassino
- Discussão: trade-off exploration-exploitation

Greedy/∈-Greedy

- Greedy: inviável e inadequado
- ∈-Greedy: exploit, mas chance de “∈” de explorar outra alavanca
- Alternativa viável, apesar de simplista, de encarar o problema
- Prática1

UCB

- “Upper-Confidence-Bound Action Selection”
- Ideia: dar uma chance a uma alavanca que você não testou muito até agora
- Expressão matemática: [explore + exploit] (balanço)
- Prática2

$$A_t = \arg\max_a \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Thompson Sampling

- Perspectiva bayesiana
- Ótimos resultados experimentais
- Começo: escolher “flat priors” sobre cada alavanca (ou seja, distribuição que indica desconhecimento)
- repetição: -amostre de cada distribuição
  -escolha a que deu maior retorno
  -atualize a distribuição com o resultado
- Prática3

Método do Gradiente

- Preferência no instante t para uma ação a -> $H_t(a)$ (Softmax)
- Política no instante t -> π_t
- Ação escolhida pela política -> A_t

$$p_a = \frac{e^{H_t(a)}}{\sum_{i=0}^{k} e^{H_t(i)}}$$

Método do Gradiente

- Caso a ação seja a que foi escolhida

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha(R_t - \bar{R}_t)\,(1 - p_{A_t})$$

- Caso a ação seja diferente da que foi escolhida

$$H_{t+1}(a) \doteq H_t(a) - \alpha(R_t - \bar{R}_t)p_a$$

- Prática4

Conclusão da aula de hoje

- Partes essenciais de um modelo de RL
- Um caso clássico de motivação ao RL (n-armed-bandit)
- Algoritmos primitivos que são predecessores dos atuais de RL

O que veremos na próxima

- Solidificaremos a base dos modelos de RL (aprofundar os pilares)
- Interações do modelo com o mundo (Markov-Decision-Process, Retorno)
- Políticas e Funções-Valor
- Técnica de calcular Função-Valor e Policy na prática (Dynamic Programming)

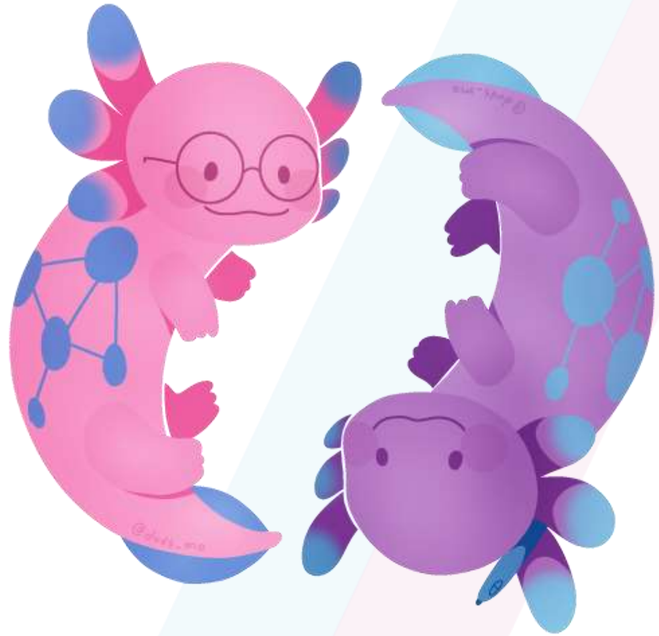
data@icmc.usp.br

@data.icmc

/c/DataICMC

/icmc-data

data.icmc.usp.br



obrigado!