



**UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO**

Projeto de Inteligência Artificial- Q-Learning

Enzo Emanuel Costa Maia
Jackson Santana Carvalho Junior
Adam Guilherme Mendes Lima
Mateus Henrique Silva de Melo
Samuel Guimarães Silva

**São Cristóvão, Sergipe
2026**

Sumário

1. Especificação do Problema	3
1.1 Introdução	3
1.2 Escopo do Projeto	3
1.3 Descrição do Problema	4
1.4 Tabela de Correspondência com o Pseudocódigo	5
2. Fundamentação Teórica	5
2.1. Q-Learning	5
2.2 Visualização e Resultados Obtidos	6
3. Conclusão	6
3.1. Considerações Finais	6
REFERÊNCIAS	7

1. Descrição do Projeto

1.1 Introdução

A área de Inteligência Artificial (IA) tem como um de seus principais objetivos o desenvolvimento de agentes capazes de aprender a partir da interação com o ambiente, dentro desse contexto, o Aprendizado por Reforço destaca-se por permitir que um agente aprenda por tentativa e erro, recebendo recompensas ou penalidades de acordo com suas ações.

Este trabalho tem como objetivo a implementação do algoritmo Q-learning, um dos métodos mais clássicos e fundamentais do Aprendizado por Reforço, aplicando-o a um problema simples e didático: a navegação de um agente em um ambiente do tipo Grid World, onde a escolha desse problema visa facilitar a compreensão do funcionamento do algoritmo, bem como permitir uma clara correspondência entre o pseudocódigo apresentado em aula e a implementação prática desenvolvida pela equipe. O foco principal do projeto é evidenciar como um agente inteligente pode aprender uma política ótima de ações a partir da interação contínua com o ambiente, sem possuir conhecimento prévio sobre sua dinâmica interna. Dessa forma, o projeto prioriza a clareza conceitual, a fidelidade ao pseudocódigo apresentado em aula e a facilidade de compreensão do processo de aprendizado.

1.2 Escopo do Projeto

O escopo deste projeto consiste no desenvolvimento de uma aplicação computacional completa, funcional e didática que demonstre o uso do algoritmo de Aprendizado por Reforço Q-learning na resolução de um problema de tomada de decisão sequencial em ambiente discreto.

O foco principal do projeto é evidenciar como um agente inteligente pode aprender uma política ótima de ações a partir da interação contínua com o ambiente, sem possuir conhecimento prévio sobre sua dinâmica interna. Dessa forma, o projeto prioriza a clareza conceitual, a fidelidade ao pseudocódigo apresentado em aula e a facilidade de compreensão do processo de aprendizado, dentro do escopo definido, o projeto contempla os seguintes tópicos:

- A implementação integral do algoritmo Q-learning, respeitando a estrutura, variáveis e equações do pseudocódigo utilizado como referência teórica;
- A modelagem de um ambiente Grid World determinístico, com estados, ações e recompensas bem definidos;
- O treinamento do agente ao longo de múltiplos episódios, permitindo a observação da evolução do aprendizado;
- A análise do comportamento do agente após o treinamento, verificando a convergência para caminhos mais eficientes;

- O desenvolvimento de uma interface de visualização que auxilie na interpretação dos resultados obtidos;
- A produção de documentação técnica detalhada, incluindo este relatório e comentários no código-fonte.

1.3 Descrição do Problema

O problema abordado neste trabalho consiste em permitir que um agente inteligente aprenda a se locomover de forma autônoma em um ambiente bidimensional representado por uma grade (Grid World), partindo de uma posição inicial fixa até alcançar um estado objetivo previamente definido. O ambiente é composto por uma matriz de dimensões 5x5, na qual cada célula representa um estado possível do agente. O estado inicial localiza-se no canto superior esquerdo da grade, enquanto o estado objetivo encontra-se no canto inferior direito, e o agente pode executar quatro ações distintas:

- Mover-se para cima
- Mover-se para baixo
- Mover-se para a esquerda
- Mover-se para a direita.

A dinâmica do ambiente é determinística, logo, dada uma ação válida, o próximo estado do agente é sempre o mesmo, e caso uma ação leve o agente para fora dos limites da grade, o agente permanece no estado atual, evitando estados inválidos. O problema caracteriza-se como um cenário de Aprendizado por Reforço, no qual o agente deve aprender, ao longo de vários episódios, quais ações são mais vantajosas em cada estado para maximizar a recompensa acumulada ao longo do tempo.

Para guiar o aprendizado do agente, foi estabelecido um sistema de recompensas e penalidades. O objetivo final (Ouro) concede uma recompensa positiva de +1.0. O mapa também possui obstáculos na forma de buracos, que geram uma penalidade de -1.0 e encerram o episódio. Além disso, para incentivar o agente a encontrar o caminho mais curto e eficiente, cada passo dado em um espaço seguro resulta em uma pequena penalidade de -0.04. Dessa forma, o agente é forçado a equilibrar a exploração cautelosa do ambiente com a necessidade de chegar ao objetivo rapidamente.

1.4 Tabela de Correspondência com o Pseudocódigo

Linha do Pseudocódigo (Imagem)	<i>persistent:</i> Q, N_{sa}, s, a	<i>if s is not null then</i>	<i>increment</i> $N_{sa}[s, a]$	$Q[s, a] \leftarrow \dots + \alpha(N_{sa}) \dots$	$\operatorname{argmax}_{a'} f(Q[s', a'], N_{sa}[s', a'])$
Linha do Código Python (app.py)	<code>self.Q = {}, self.N_sa = {}, self.s = None (no __init__)</code>	<code>if self.s is not None:</code>	<code>self.N_sa[(self.s, self.a)] = current_n + 1</code>	<code>self.Q[(self.s, self.a)] = ...</code>	Método $f(u, n)$ e loop de decisão

- Inicialização de dicionários para armazenar os valores de ação (Q) e a frequência de visitas (N) aos pares estado-ação.
- Verifica se não é o primeiro passo do agente no ambiente para garantir que existe um estado anterior a ser atualizado.
- Atualiza a contagem de quantas vezes o agente executou a ação A no estado S.
- Atualiza o valor Q utilizando uma taxa de aprendizado alpha decrescente, baseada no número de visitas ($1 / N$).
- Implementa a função de exploração otimista: se um estado foi pouco visitado ($N < N_e$), o agente é forçado a explorá-lo.

2. Fundamentação Teórica

2.1 Q-Learning

O Q-Learning é um algoritmo de Aprendizado por Reforço model-free, o que significa que o agente não precisa conhecer as probabilidades de transição do ambiente para aprender a política ótima. O algoritmo baseia-se na atualização interativa dos chamados Q-values, que representam a utilidade esperada de se tomar uma determinada ação em um estado específico.

A atualização ocorre com base na Equação de Bellman, no entanto, o diferencial desta implementação, é o tratamento da dicotomia entre exploração e exploração, onde ao invés de utilizar abordagens aleatórias, o algoritmo implementado utiliza uma Função de Exploração Otimista, combinada com uma tabela de frequências, logo, retorna uma recompensa otimista, e é isso que garante que o agente explore sistematicamente todas as regiões desconhecidas do Grid World antes de convergir para o caminho ótimo.

2.2 Visualização e Resultados Obtidos

Para atender à proposta visual e interativa do projeto, a interface gráfica foi desenvolvida utilizando a biblioteca Streamlit, onde a interface não apenas renderiza o tabuleiro e o movimento do agente de forma lúdica, mas também expõe os hiperparâmetros (número de episódios, velocidade) para o usuário. Durante os testes, observou-se que em episódios iniciais o agente cai frequentemente nos buracos ou fica preso em ciclos devido à exploração forçada. Contudo, após cerca de 100 a 300 episódios, os valores da tabela Q convergem, e o agente demonstra um comportamento puramente exploratório, navegando diretamente do ponto de partida até o ouro pelo caminho mais curto, comprovando a eficácia da implementação.



3. Conclusão

3.1 Conclusões finais

O projeto proposto atingiu com êxito todos os objetivos propostos, sendo pautado na implementação do algoritmo Q-Learning a partir do pseudocódigo de referência, tal atividade nos permitiu compreender, na prática, os desafios do Aprendizado por Reforço, especialmente o equilíbrio entre explorar o desconhecido e maximizar recompensas, e isso resultou em uma aplicação não apenas funcional e aderente aos requisitos teóricos, mas também altamente didática, pois o agente provou ser capaz de iniciar, sem qualquer conhecimento prévio do Grid World, puramente através da interação estocástica e da atualização da sua Tabela Q, encontrar consistentemente a política ótima de navegação.

REFERÊNCIAS

RUSSELL, S. J.; NORVIG, P. Inteligência Artificial: Uma Abordagem Moderna. 3. ed. Rio de Janeiro: Elsevier, 2013.

WATKINS, C. J. C. H.; DAYAN, P. Q-learning. Machine Learning, v. 8, n. 3-4, p. 279-292, 1992. Disponível em: <https://link.springer.com/article/10.1007/BF00992698>. Acesso em: 19 fev. 2026.

STREAMLIT. Streamlit Documentation. 2026. Disponível em: <https://docs.streamlit.io/>. Acesso em: 19 fev. 2026.