Estudo de aplicação em Reinforcement Learning

Quantum Finance - **Turma 9DTSR**

Nome: Jéssica Portela de Castro RM: 359735

Nome: Tiago Freire Barbosa RM: 358404

Introdução e Problemática

No contexto atual dos mercados financeiros, a tomada de decisões eficazes sobre a compra, venda ou manutenção de ativos financeiros é crítica para a obtenção de retornos positivos e gerenciamento eficiente de riscos. A crescente complexidade e volatilidade dos mercados têm impulsionado o interesse em soluções tecnológicas avançadas, especialmente aquelas que envolvem inteligência artificial.

Dentro desse campo, o aprendizado por reforço destaca-se por sua capacidade de aprender políticas ótimas por meio da interação direta com o ambiente, adaptando-se às condições dinâmicas do mercado.

Assim, a **QuantumFinance** visa explorar o potencial do RL aplicado especificamente aos ativos **Vale**, **Petrobras** e **Brasil Foods** para desenvolver um fundo automatizado capaz de realizar operações financeiras baseadas em análises históricas e adaptativas.

Motivação e Objetivo

A motivação deste trabalho reside na necessidade de automação e aprimoramento contínuo das decisões financeiras, reduzindo interferências subjetivas e emocionais inerentes ao processo humano de negociação.

O objetivo central é desenvolver um agente baseado em Reinforcement Learning, utilizando técnicas avançadas como Deep Q-Networks (DQN), que seja capaz de realizar decisões eficientes de negociação: comprar, vender ou manter ativos com base em dados históricos dos preços das ações selecionadas.

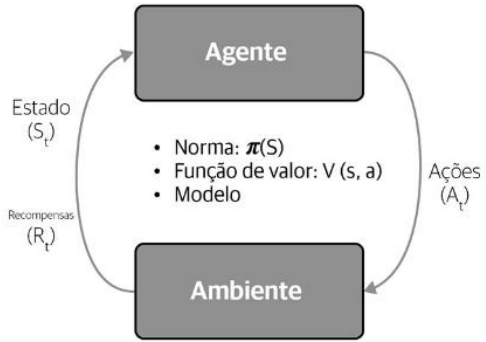
Além disso, pretende-se avaliar a eficácia desse agente utilizando métricas financeiras reconhecidas, como retorno total, índice de Sharpe e máximo drawdown.

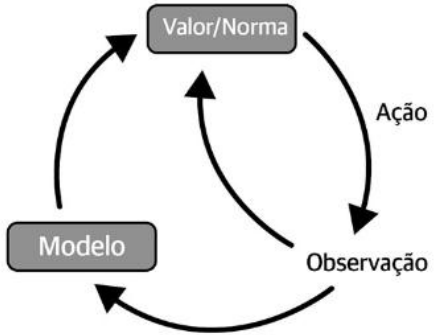
Apresentação da Aplicação estudada / Estrutura

Definição do Problema de RL

A aplicação do aprendizado por reforço ao mercado financeiro requer uma definição detalhada e clara do problema, identificando três componentes fundamentais: estados, ações e recompensas.

O objetivo final do agente de RL é encontrar uma política ótima, ou seja, um conjunto de ações que maximize o retorno acumulado (recompensa) ao longo de um período de tempo especificado. O agente busca alcançar um equilíbrio eficaz entre risco e retorno, proporcionando uma performance consistente e lucrativa.





Estados

O estado em um problema de RL é uma representação da situação atual do ambiente. Neste contexto financeiro específico, o estado pode ser definido pelos seguintes elementos:

Preços atuais dos ativos: os valores de mercado das ações Vale, Petrobras e Brasil Foods em cada momento do tempo.

Indicadores técnicos: métricas quantitativas utilizadas para analisar tendências e desempenho das ações, como médias móveis, índices de força relativa (RSI) e volatilidade histórica.

Posição atual do portfólio: quantidade de ações de cada ativo atualmente detidas pelo agente.

Saldo disponível: quantidade de dinheiro em caixa para ser investida.

Esses componentes juntos fornecem ao agente uma visão completa e atualizada, permitindo decisões informadas e adequadas.

Ações

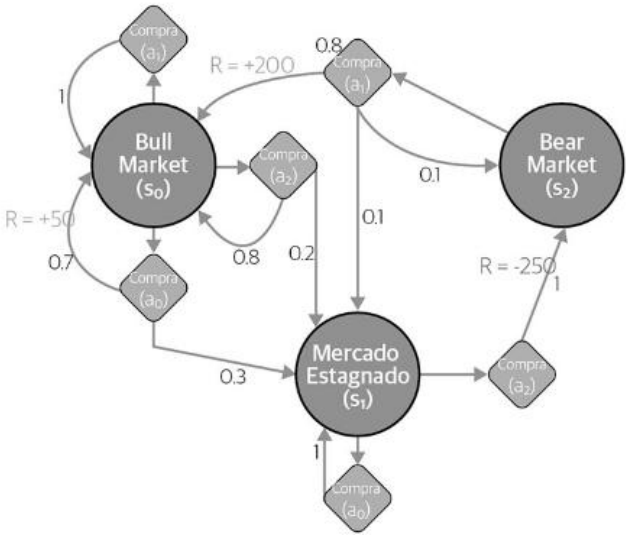
As ações são as decisões disponíveis para o agente realizar a cada passo. Neste problema, as ações são discretas e específicas para cada ativo, limitando-se a:

Comprar: adquirir uma unidade adicional do ativo, desde que haja saldo disponível suficiente.

Vender: desfazer-se de uma unidade do ativo que já esteja na carteira, desde que exista posição no ativo.

Manter: não realizar nenhuma alteração na posição do ativo em questão.

Essas ações simplificadas permitem ao agente realizar operações pontuais e frequentes, ajustando dinamicamente o portfólio.



Recompensas

A recompensa é um retorno quantitativo recebido pelo agente após a realização de uma ação, indicando o quanto essa ação foi benéfica ou prejudicial ao objetivo de maximizar retornos. No contexto financeiro, a recompensa é representada pelo lucro ou prejuízo obtido após cada ação, especificamente pela variação do patrimônio total entre passos consecutivos da simulação.

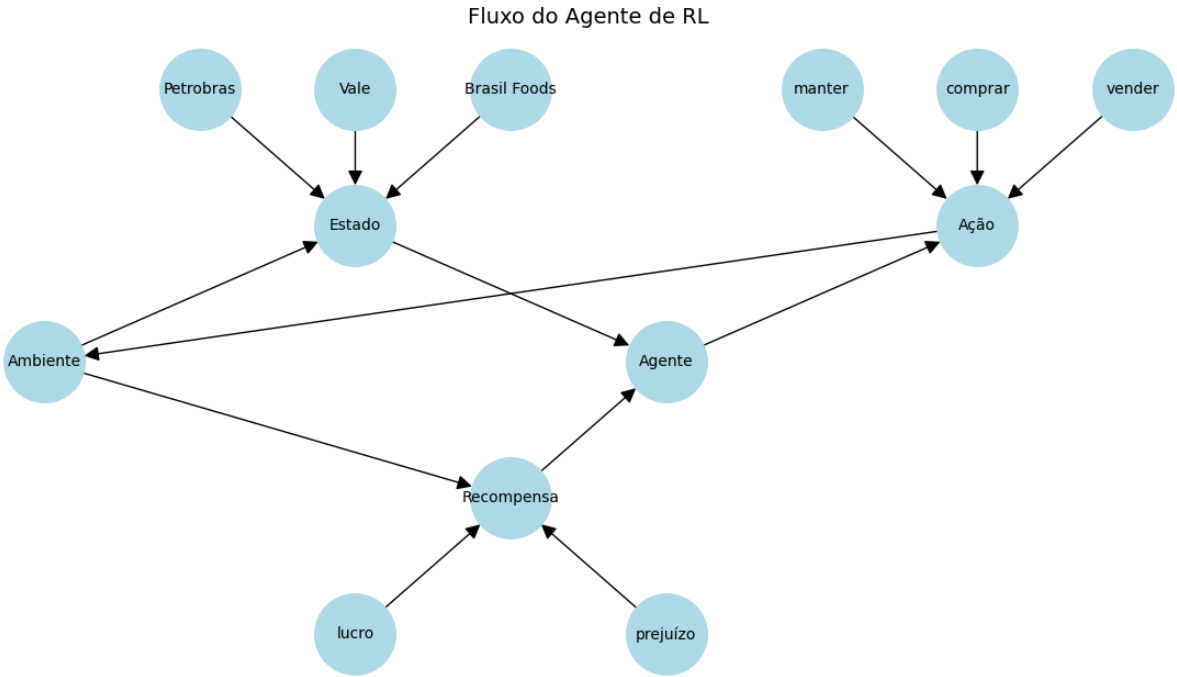
Matematicamente, a recompensa é definida como:

$$\text{Recompensa} = \text{Patrimônio}_{t+1} - \text{Patrimônio}_{t}$$

onde:

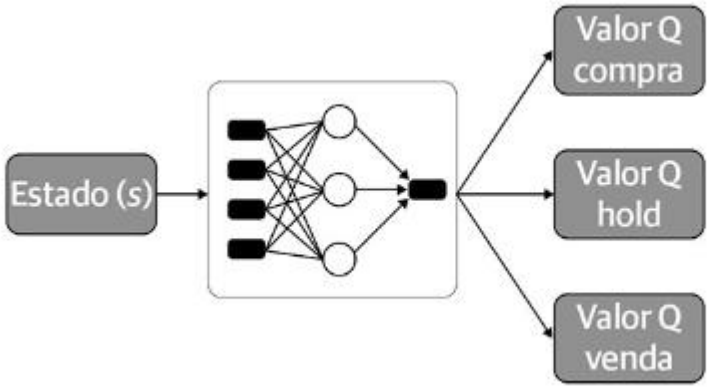
- Patrimônio_{t} é o valor total do portfólio (ações e saldo disponível) no passo atual.
- Patrimônio_{t+1} é o valor total do portfólio após a ação do agente.

Desenho do Agente



Implementação do Agente RL usando DQN

O Deep Q-Network (DQN) é um método de Reinforcement Learning que combina aprendizado profundo com o tradicional Q-Learning. O objetivo principal é treinar o agente para escolher ações que maximizam sua recompensa futura.



Coleta e Pré-Processamento dos Dados

Para este projeto, utilizaremos a biblioteca **yfinance**, que oferece acesso gratuito e conveniente aos preços ajustados (considerando dividendos e splits) dos ativos Vale (VALE3.SA), Petrobras (PETR4.SA) e Brasil Foods (BRFS3.SA)

O período utilizado será referente a 01-01-2022 a 31-12-2024

Em seguida utilizamos um método do pandas que preenche valores faltantes com o valor válido mais recente disponível ("forward fill"), selecionamos apenas a coluna "preço de fechamento ajustado" (Adj Close), e removemos qualquer valor restante faltante com dropna

Tabela obtida da API YFinance

Ticker	BRFS3.SA	PETR4.SA	VALE3.SA
Date			
2022-01-03	22.007359	11.173077	59.548798
2022-01-04	21.249138	11.215325	58.846432
2022-01-05	21.514515	10.781305	59.403744
2022-01-06	23.030956	10.773625	60.602356
2022-01-07	23.315289	10.823557	64.129478

Definição do Ambiente

Estado:

- Preços atuais das ações
- Posições do agente em cada ação
- Saldo disponível

Ações:

- Comprar uma unidade (se houver saldo suficiente)
- Vender uma unidade (se possuir a ação)
- Manter (não fazer nada)

Recompensa:

A recompensa é o lucro ou prejuízo obtido pelo agente entre dois passos consecutivos

Treinamento do Agente DQN

Neste passo, utilizamos o algoritmo DQN do pacote stable_baselines3, um framework que simplifica o uso de técnicas avançadas de RL

O DQN treina uma rede neural para estimar o valor esperado das ações (Q-value):

Entrada da rede: o estado atual (preços, posições, saldo)

Saída da rede: valor esperado para cada ação possível

MlpPolicy: é a rede neural padrão (multilayer perceptron)

total_timesteps: indica por quantos passos de simulação o modelo será treinado.

Durante o treinamento, o DQN explora diferentes ações e aprende quais decisões levam a maiores recompensas

Após a realização do treinamento do agente DQN é gerado um log automaticamente pela biblioteca Stable-Baselines3. Ela resume o desempenho e o progresso do agente em termos de exploração, recompensa, tempo de treinamento e aprendizado.

Interpretação do Log de Treinamento do Agente DQN

Interpretação do Log de Treinamento do Agente DQN

Categoria	Métrica	Significado
rollout/	ep_len_mean	Comprimento médio dos episódios (número de passos por episódio)
	ep_rew_mean	Recompensa média por episódio Mede o quão bem o agente está se saindo
	exploration_rate	Taxa de exploração atual (epsilon). Indica a frequência de ações aleatórias
time/	episodes	Quantidade de episódios completados durante o treinamento
	fps	Passos por segundo (velocidade do treinamento)
	time_elapsed	Tempo total decorrido desde o início do treinamento
	total_timesteps	Número total de passos de simulação já realizados
train/	learning_rate	Taxa de aprendizado da rede neural durante o treinamento
	loss	Valor médio da função de perda, que indica o erro do modelo ao prever ações
	n_updates	Número de atualizações dos pesos da rede neural

```
---------------------------------
| rollout/            |         |
|    ep_len_mean      | 748     |
|    ep_rew_mean      | 1.07e+03 |
|    exploration_rate | 0.05    |
| time/               |         |
|    episodes         | 64      |
|    fps              | 485     |
|    time_elapsed     | 98      |
|    total_timesteps  | 47872   |
| train/              |         |
|    learning_rate    | 0.0001  |
|    loss             | 28      |
|    n_updates        | 11942   |
---------------------------------
```

Interpretação do Log de Treinamento do Agente DQN

O agente está sobrevivendo bastante nos episódios e obtendo recompensas positivas.

O treinamento está quase concluído (47.872 de 50.000 timesteps).

A exploração foi reduzida para 5%, o que indica que o agente está quase só explorando sua política aprendida.

A perda ainda está alta (28), sugerindo que a rede pode não estar perfeitamente estável. Isso pode melhorar com ajustes no modelo, mais dados ou tuning de hiperparâmetros.

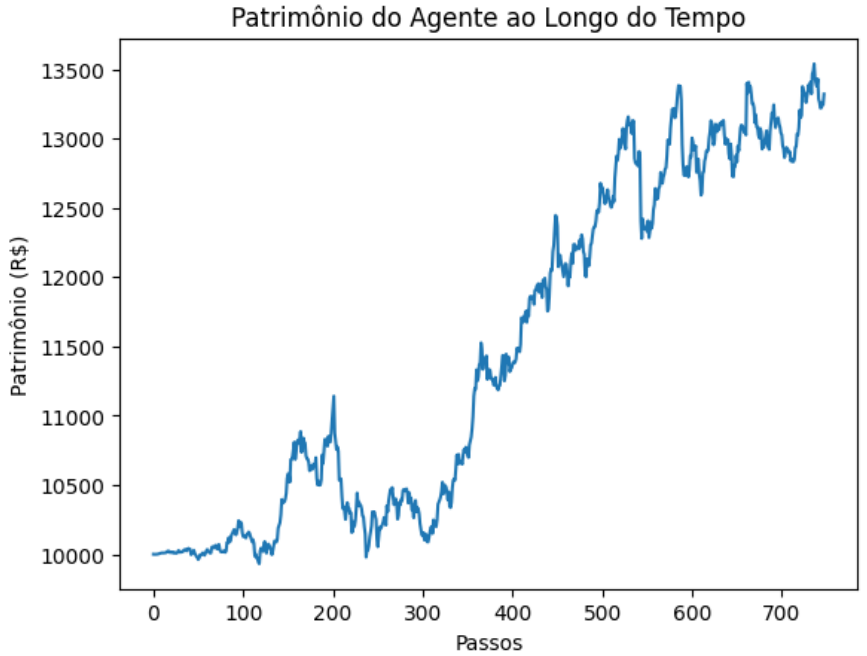
Simulação das Operações

- O agente observa o estado atual (preços, posições, saldo)
- Escolhe uma ação com base na política aprendida (modelo treinado)
- O ambiente realiza a ação e atualiza saldo e posições
- Registra-se a evolução do patrimônio

O agente partiu de um valor inicial R$ 10.000 e chegou a mais de R$ 13.500, o que representa um retorno total positivo.

O comportamento da curva mostra que o modelo não está apenas acumulando aleatoriamente, mas aprendendo padrões que levam a decisões eficazes.

Pode-se dizer que o agente teve sucesso em seu treinamento, dado que ele aumentou o patrimônio de forma progressiva e sustentada.



Avaliação do Desempenho

As seguintes métricas financeiras são utilizadas para avaliar o desempenho do agente de Reinforcement Learning:

Retorno Total

$$\text{Retorno Total} = \frac{P_{\text{final}} - P_{\text{inicial}}}{P_{\text{inicial}}}$$

Onde:

- P_inicial: patrimônio no início da simulação
- P_final: patrimônio no final da simulação

Sharpe Ratio (Anualizado)

$$\text{Sharpe Ratio} = \frac{\mu_r}{\sigma_r} \times \sqrt{252}$$

Onde:

- μ_r: média dos retornos diários
- σ_r: desvio padrão dos retornos diários
- 252: número aproximado de dias úteis no ano

Máximo Drawdown (queda máxima do patrimônio)

$$\text{Máximo Drawdown} = \min\left(\frac{P_t}{P_{\text{máximo}}} - 1\right)$$

Onde:

- P_t: patrimônio no tempo t
- P_máximo: maior patrimônio observado até o tempo t

```
Sharpe Ratio: 1.14
Retorno Total: 33.21%
Máximo Drawdown: -3.48%
```

Avaliação do Desempenho

- **Retorno Total: 33.21%**
  - O agente aumentou seu patrimônio em 33,21% ao longo da simulação.
  - Isso significa que, partindo de 10.000, por exemplo, ele terminou com aproximadamente 13.321.
  - **Conclusão:** um excelente resultado de rentabilidade bruta em um único ciclo de treinamento.

- **Sharpe Ratio: 1.14**
  - Mede o retorno ajustado ao risco.
  - Um Sharpe Ratio acima de 1.0 indica que o agente está gerando bons retornos em relação ao risco que está correndo.
  - Entre 1 e 2: bom, acima de 2: excelente
  - **Conclusão:** o agente está gerando lucros de maneira eficiente em termos de risco, com baixo nível de volatilidade relativa.

- **Máximo Drawdown: -3.48%**
  - Representa a maior queda do patrimônio desde um topo anterior até um fundo posterior, durante o período analisado.
  - Uma queda máxima de apenas 3,48% é muito baixa e indica que o agente evitou grandes perdas.
  - **Conclusão:** o agente teve ótimo controle de risco, enfrentando variações mínimas mesmo em momentos desfavoráveis.

Considerações e Potencial

A aplicação do aprendizado por reforço no domínio financeiro apresenta grande potencial para transformar a maneira como fundos de investimento operam, oferecendo decisões automatizadas baseadas em aprendizado contínuo e adaptativo.

Contudo, é fundamental reconhecer os desafios existentes, tais como o risco de overfitting aos dados históricos, a necessidade de validação robusta e o gerenciamento prudente dos riscos inerentes à automação financeira.

Se bem implementado, o agente pode não apenas proporcionar uma gestão mais eficiente dos investimentos, como também oferecer insights estratégicos valiosos para investidores e gestores financeiros, destacando-se como um instrumento inovador e competitivo no mercado financeiro contemporâneo.

Referências

TATSAT, Hariom; PURI, Sahil; LOOKABAUGH, Brad. *Blueprints de aprendizado de máquina e ciência de dados para finanças: desenvolvendo desde estratégias de trades até robôs advisors com Python*. Tradução: Alta Books. Rio de Janeiro: Alta Books, 2021.