O aprendizado por reforço (Q-Learning) consegue gerenciar portfólios no mercado de criptomoedas?

Ludmilla Mattos
Rafael Morais
Laerte Takeuti

Universidade de Brasília

05 de outubro de 2018

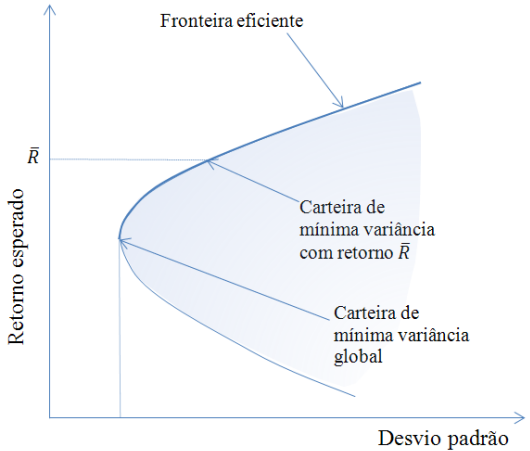
Motivação

- Alocação de portfólios
- Cryptomoedas
- Q-Learning

Alocação de Portfólios

- Carteira com diversos ativos
- Diversificar para reduzir o risco e maximizar o retorno
- Principio do investidor racional
- Markowitz (1952)

Alocação de Portfólios



Cryptomoedas

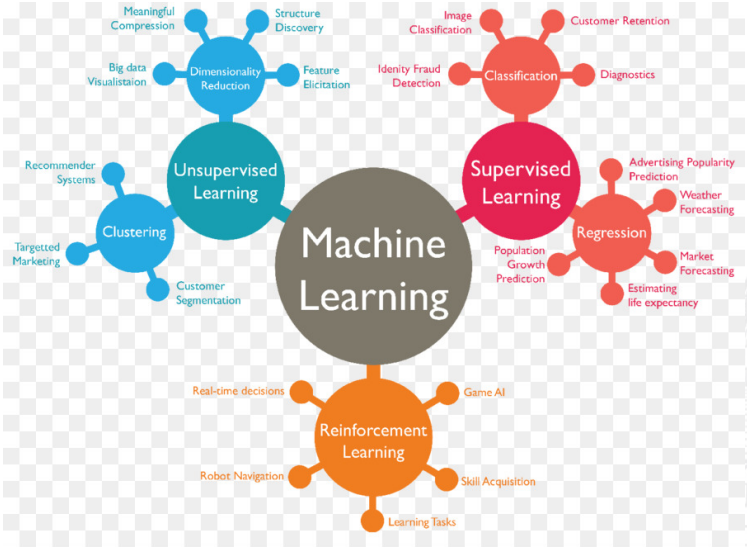
- Bitcoin surge em 2009, autoria atribuida a Satoshi Nakamoto
- Tecnologia de Blockchain
- Mais de 1900 moedas ativas [1]

[1] https://www.coinlore.com/all_coins

Q-Learning

- Aprendizado por Reforço
- Processo markoviano de decisão
- Free-model policy

Q-Learning


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Processo Markoviano de Decisão

Técnica capaz de lidar com problemas que envolvem sequências de decisões orientadas a um objetivo. Deseja-se obter uma política ótima na qual o agente recebe o máximo de retorno médio.

- Conjunto Finito de **Estados**
- **Ações** possíveis para cada estado
- Probabilidades de transições entre os estados
- Função **Retorno** entre os estados dado a ação

$$V_s^{\pi} = R_s(\pi(s)) + \gamma \sum_{s'} P_{ss'}[\pi(s)] V^{\pi}(s')$$

Model-Free policy

Não exige o conhecimento das probabilidades de transição.

$$Q^{Novo}(s_t, a_t) = (1-\alpha)\ \underbrace{Q(s_t, a_t)}_{Valor\ Antigo} + \underbrace{\alpha}_{Tx\ Aprend.}\ \overbrace{[\ \underbrace{r_t}_{Recompensa} + \underbrace{\gamma}_{Fator\ Desconto}\ \underbrace{\max_a Q(s_{t+1}, a)}_{Est.\ do\ valor\ futuro}\ ]}^{Valor\ Aprendido}$$

- observa o estado s_t
- seleciona e executa uma ação a_t
- observa o estado subsequente s_{t+1}
- recebe uma recompensa imediata r_t
- ajusta Q^{Novo}

Estratégia de Referência

Baseada na intuição do passeio aleatório (*Ramdon Walk - RW*), isto é, toda informação do momento presente encontra-se no instante anterior. Para cada instante de tempo considera-se o retorno anterior de todas as moedas para então alocar 100% do portfólio na que apresentou o melhor desempenho.

Teste de habilidade de predição superior

Sua hipótese nula é a de que nenhuma predição considerada é superior à referência, isto é, se rejeitada a hipótese inferimos que ao menos uma das séries preditas é superior.

Estados: são estabelecidos com base na trajetória do preço das moedas. Foi considerado a combinação do sinal obtido do retorno de cada moeda.

Tabela 1: Estados Possíveis

S	1	2	3	$\cdots$	7	8
Bitcoin	+	+	+	$\cdots$	-	-
Ethereum	+	+	-	$\cdots$	-	-
Litcoin	+	-	+	$\cdots$	+	-

Ações: considera o montante alocado a cada moeda, dado por valores entre 0,0 e 1.

Tabela 2: Alocação do Portfólio (%)

A	1	2	3	$\cdots$	60	61	62
Bitcoin	100	90	80	$\cdots$	0	0	0
Ethereum	0	10	20	$\cdots$	20	10	0
Litcoin	0	0	0	$\cdots$	80	90	100

Recompensas: devem representar o prêmio imediato acarretado pela ação. Uma vez que a ação estabelece quanto do capital é alocado, a recompensa é dada pela soma dos retornos ponderada pela alocação:

$$r_t = A_t * R_t$$

Sendo,

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$

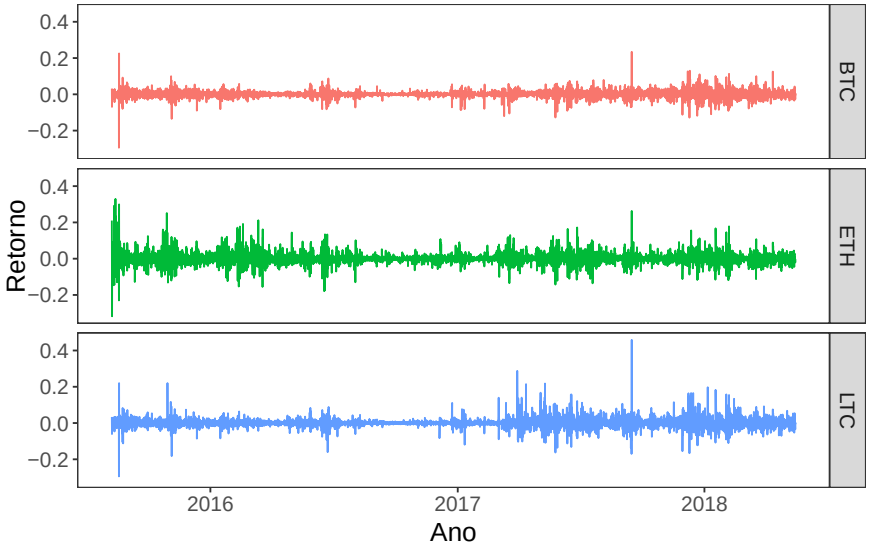
Matriz Q

Matriz de Estados e Ações de dimensão $[8 \times 62]$

Tabela 3: Alocação do Portfólio (%)

	(100,0,0)	(90,10,0)	(80,20,0)	$\cdots$	(0,10,90)	(0,0,100)
+ + +						
+ + -						
$\vdots$						
+ - -						
- - -						

Dados



Dados

Os dados foram coletados no dia 4 de abril de 2018 de forma automática do site https://poloniex.com. Foram utilizados 6068 pontos de dados referentes ao preço de mercado das moedas obtidos com intervalos de três horas desde agosto de 2015 até maio de 2018. A base de dados foi dividida em fases de treinamento, validação e teste:

Tabela 4: Partição da base de dados

Partição	Período	Tamanho
Treinamento	Ago-2015 $\sim$ Mai-2017	3.900
Teste	Mai-2017 $\sim$ Mai-2018	2.168

Resultados

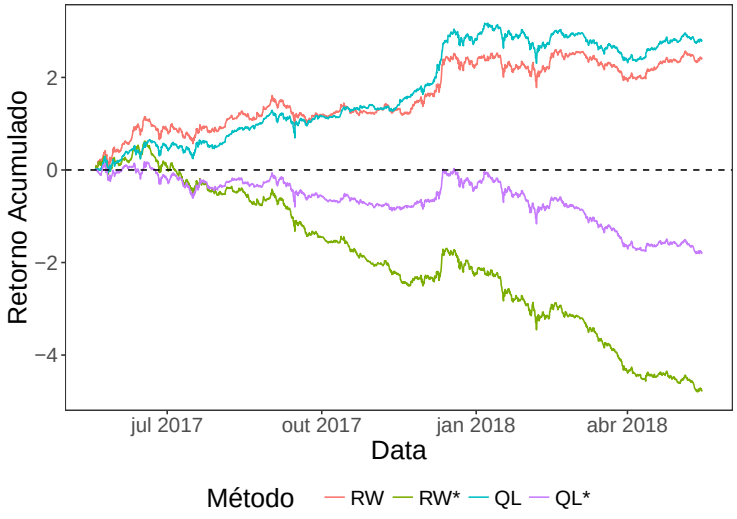

Retorno Acumulado
2
0
−2
−4
jul 2017
out 2017
jan 2018
abr 2018
Data
Método
RW
RW*
QL
QL*

Tabela 5: Métricas de desempenho e teste SPA de comparação para as Estratégias

Estratégia	Retorno Médio	Volatilidade	*Sharpe Ratio*	teste SPA
RW	1,61	0,45	2,41	
QL	1,74	0,99	1,75	0,298
RW*	-1,96	2,29	-1,30	
QL*	-0,63	0,26	-1,25	$<$ 0,001

Conclusões

- Explorar Espaços mais informativos
- Incorporar a volatilidade no cálculo de recompensa

Código-Fonte

https://github.com/ludmattos/qLearningFinance