BLACKJACK

REINFORCEMENT LEARNING AGENT

GRUPO F: André Pereira, Miguel Amaro, ZhiXu Ni

Introdução a Sistemas Inteligentes e Autonomos, 2023/2024

REGRAS

INÍCIO

O jogador começa com 2 cartas viradas para cima, enquanto que o dealer começa com 1 carta virada para cima e uma para baixo.

OBJETIVO

Tanto o jogador como o dealer têm o objetivo de aproximar o mais possível, ou igualar a sua pontuação a 21, sem nunca ultrapassar esse valor.

PONTUAÇÃO

A qualquer ponto do jogo, a pontuação do jogador (e do dealer) corresponde à soma do valor das cartas na sua posse.

AÇÕES

O jogador pode escolher pedir mais cartas quando quiser (hit), até ao momento que ultrapassa uma pontuação de 21 e perde (bust). Se tiver uma pontuação inferior ou igual a 21, o jogador pode escolher ficar com a pontuação atual (stick).

VALORES

Cartas de figuras valem 10 pontos; cartas de número valem os mesmos pontos que o seu número: o às vale 1 ou 11 pontos (de acordo com a escolha do jogador).

FIM

Se o jogador ultrapassa uma pontuação de 21, perde. Se escolher stick, o dealer revela a sua segunda carta e, se tiver uma pontuação inferior a 17, retira mais cartas até atingir uma pontuação igual ou superior a 17. O vencedor é quem tiver a maior pontuação que não ultrapasse 21.

AMBIENTE ORIGINAL

AÇÕES

Neste ambiente apenas é permitido escolher entre "hit" (1) ou "stick" (0).

REWARDS

Vitórias são premiadas com 1 ponto, enquanto que derrotas são penalizadas com 1 ponto.

PARTICULARIDADES

Ações mais circunstanciais, como "splits", desistências e seguros foram excluídos (menos possibilidade de estratégias).

As cartas são retiradas de um baralho infinito com reposição (impossibilidade de contar cartas).

Estas particularidades tornam esta implementação do jogo muito desafiante!

ESTADOS

Representados por 3 valores discretos: pontuação do jogador (32 valores possíveis), potnuação do dealer (11 valores possíveis) e existência de um às utilizável na mão do jogador (2 valores possíveis).

Total de 704 possíveis estados.

(DES)VANTAGENS

Muitas poucas penalizações, exigindo pouco dos agentes treinados nestes ambientes.

Em contrapartida, o espaço de estados de tamanho reduzido (704), aliado a uma espaço de ações limitado (2), pode levar a uma rápida convergência no treino.

O NOSSO AMBIENTE

BAIXA PONTUAÇAO

Se o jogador terminar o jogo com uma pontuação demasiado baixa (<12) sofre uma penalização (-0.5), prevenindo jogadas "cobardes".

ZONA DE SEGURANÇA

Se o jogador terminar o jogo com uma pontuação geralmente vantajosa (>18 e <21), é premiado (+0.5), recompensando jogadas moderadas e estratégicas.

ALTA PONTUAÇÃO

Se o jogador terminar o jogo com uma pontuação demasiado elevada (==21) sofre uma penalizaçao (-0.5), prevenindo jogadas gananciosas.

"HITS" CONTABILIZADOS

Demasiados "hits" refletem precipitação e ganância. Jogadas com mais de 4 "hits" são penalizadas (-0.5).

ALGORITMOS UTILIZADOS

DEEP Q-NETWORK

Utiliza deep learning para aproximar a Q-value function.

PROXIMAL POLICY OPTIMIZATION

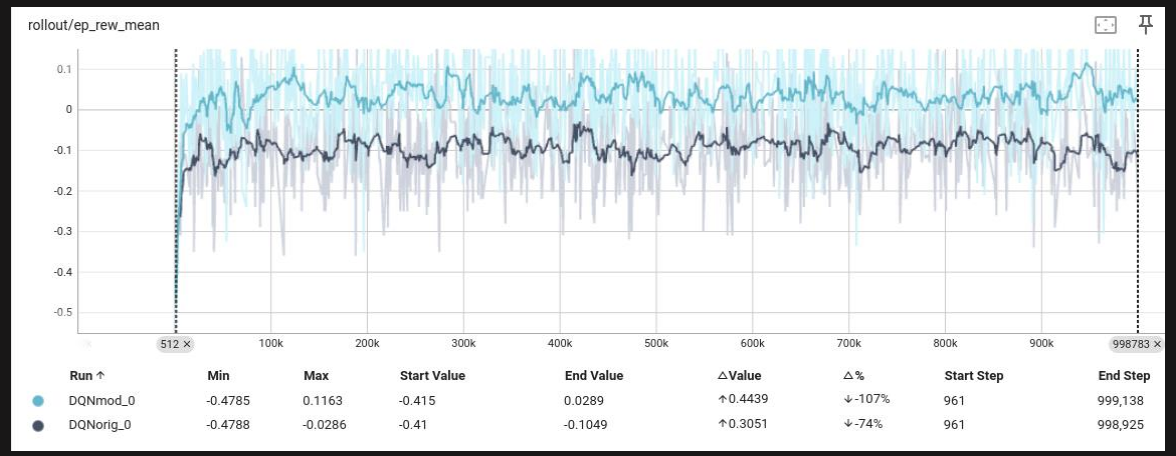
Faz pequenas atualizações na política do agente, garantindo a não perturbação do treino.

HYPERTUNNING?

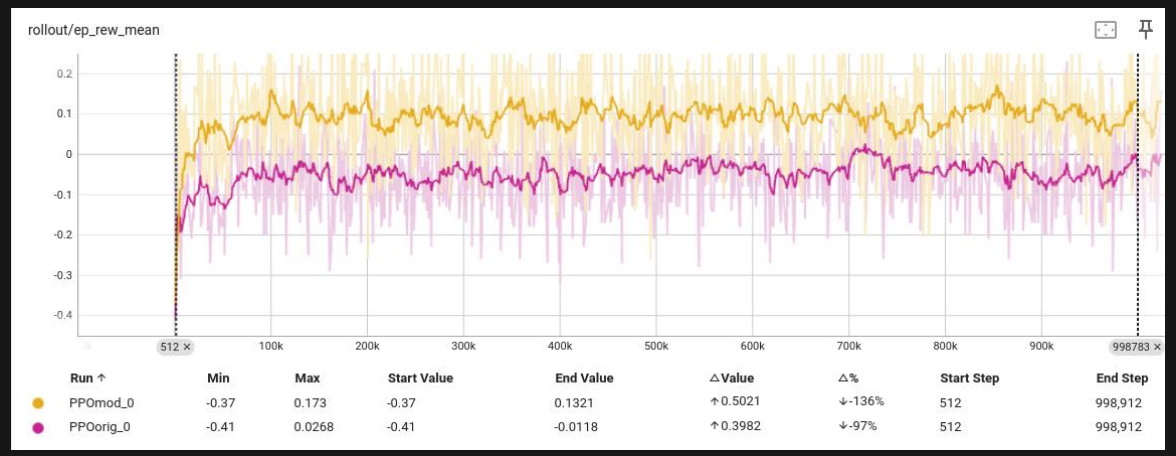
Feito através de testes de diferentes valores recomendados e iterativamente modificados. Sem muitas diferenças nos resultados finais!

TREINO

POR ALGORITMO:

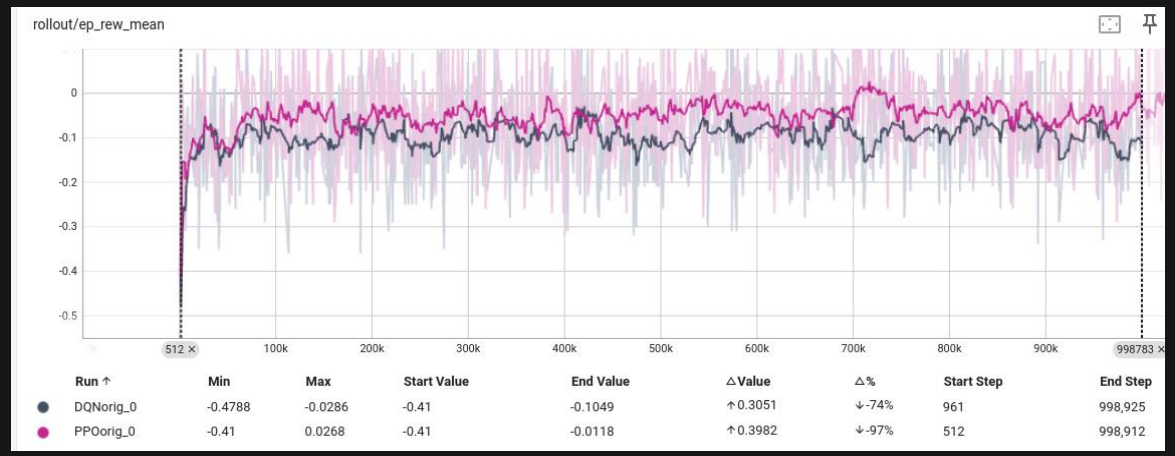


Run	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Step
DQNmod_0	-0.4785	0.1163	-0.415	0.0289	↑0.4439	↓-107%	961	999,138
DQNorig_0	-0.4788	-0.0286	-0.41	-0.1049	↑0.3051	↓-74%	961	998,925

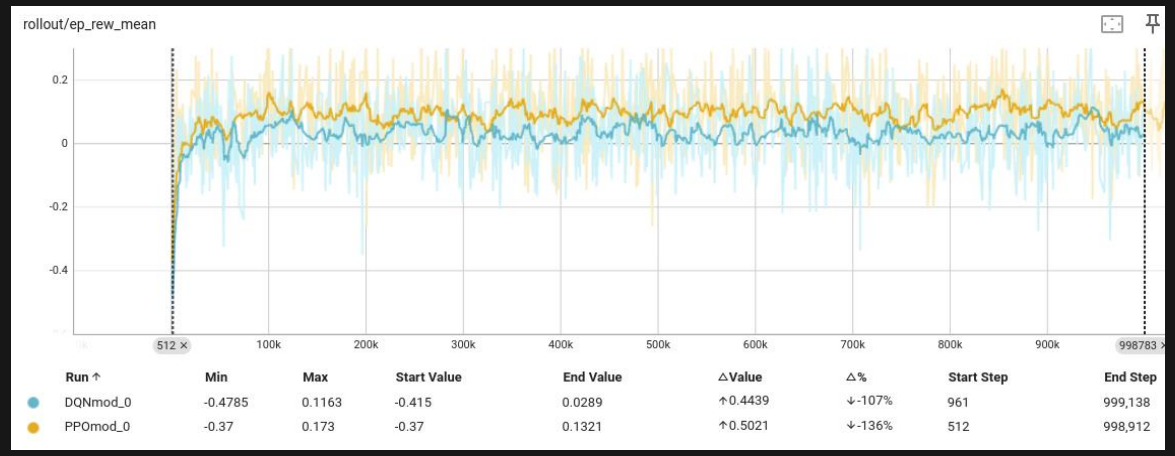


Run	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Step
PPOmod_0	-0.37	0.173	-0.37	0.1321	↑0.5021	↓-136%	512	998,912
PPOorig_0	-0.41	0.0268	-0.41	-0.0118	↑0.3982	↓-97%	512	998,912

POR AMBIENTE:

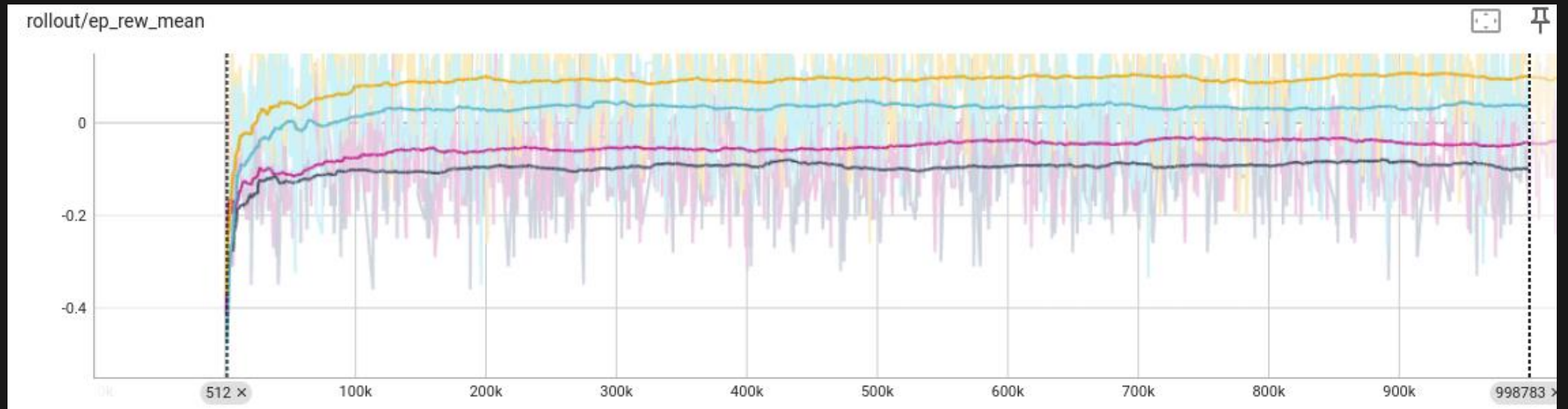


Run	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Step
DQNorig_0	-0.4788	-0.0286	-0.41	-0.1049	↑0.3051	↓-74%	961	998,925
PPOorig_0	-0.41	0.0268	-0.41	-0.0118	↑0.3982	↓-97%	512	998,912



Run	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Step
DQNmod_0	-0.4785	0.1163	-0.415	0.0289	↑0.4439	↓-107%	961	999,138
PPOmod_0	-0.37	0.173	-0.37	0.1321	↑0.5021	↓-136%	512	998,912

TREINO



Run	Min	Max	Start Value	End Value	△Value	△%	Start Step	End Step
DQNmod_0	-0.4753	0.0477	-0.415	0.0371	↑0.4521	↓-109%	961	999,138
DQNorig_0	-0.4753	-0.0788	-0.41	-0.0995	↑0.3105	↓-76%	961	998,925
PPOmod_0	-0.37	0.1077	-0.37	0.1011	↑0.4711	↓-127%	512	998,912
PPOorig_0	-0.41	-0.0308	-0.41	-0.0426	↑0.3674	↓-90%	512	998,912

TESTE

AMBIENTE ORIGINAL:

	DQN_ORIG	DQN_NMOD	PPO_ORIG	PPO_MOD
Vitória (%)	40.37	40.79	42.77	42.75
Perdas (%)	50.26	50.02	47.61	47.93
Empates (%)	9.38	9.19	9.62	9.32
Mean Reward	-0.1	-0.09	-0.05	-0.05
Mean Reward Standard Deviation	0.95	0.95	0.95	0.95

AMBIENTE MODIFICADO:

	DQN_ORIG	DQN_NMOD	PPO_ORIG	PPO_MOD
Vitória (%)	46.30	46.55	48.94	48.88
Perdas (%)	51.69	51.51	49.00	49.19
Empates (%)	2.02	1.93	2.06	1.93
Mean Reward	0.03	0.04	0.1	0.09
Mean Reward Standard Deviation	1.03	1.03	1.03	1.03

TESTE

	DQN_ORIG	DQN_NMOD	PPO_ORIG	PPO_MOD
Vitória (%)	40.37	46.55	42.76	48.88
Perdas (%)	50.26	51.51	47.61	49.19
Empates (%)	9.37	1.93	9.62	1.93
Mean Reward	-0.1	0.04	-0.05	0.09
Mean Reward Standard Deviation	0.095	1.03	0.095	1.03

O QUE SIGNIFICA ISTO?

- 35%-39% - jogador novato (jogadas aleatórias);
- 40%-44% - jogador pouco experiente (com estratégia muito básica);
- 45%-49% - jogador muito experiente (com estratégia robustas, de acordo com tabela);
- 50%-52% - jogador profissional (contador de cartas).

CONCLUSÃO

Mesmo com bons resultados, BlackJack não deixa de ser um jogo de "azar", feito para o jogador perder a longo termo.

Em futuros trabalhos, implementar um algoritmo num ambiente com estados contínuos, e com menos fatores aleatórios!