FGV EMAp

João Pedro Jerônimo

Reinforcement Learning

Exercícios do Livro

Rio de Janeiro

2025

Conteúdo

Introduction

Exercício 1.1 - *Self-Play*: Suponha que, em vez de jogar contra um oponente aleatório, o algoritmo de aprendizado por reforço descrito acima jogasse contra si mesmo, com ambos os lados aprendendo. O que você acha que aconteceria nesse caso? Ele aprenderia uma política diferente para a seleção de jogadas?

Resolução: *Como ambos os modelos são treinados para melhorar ao seu oponente, vai chegar um ponto que os jogos sempre darão empate, pois os modelos estarão bem treinados o suficiente para que a única forma de maximizar a recompensa seja ambos empatando*

Exercício 1.2 - *Symmetries*: Muitas posições do jogo-da-velha parecem diferentes, mas na verdade são iguais por causa das simetrias. Como poderíamos alterar o processo de aprendizado descrito acima para tirar proveito disso? De que maneiras essa mudança melhoraria o processo de aprendizado? Agora pense novamente: suponha que o oponente não tirasse proveito das simetrias. Nesse caso, nós deveríamos aproveitar? É verdade, então, que posições simetricamente equivalentes devem necessariamente ter o mesmo valor?

Resolução: *Poderíamos restringir a quantidade de jogadas possíveis de forma a analisar os estados simétricos como os mesmos. Isso agiliza bastante o processo de treinamento do modelo, tendo em vista que a quantidade de casos a ser analisados diminuem e muito. Sim, deveríamos tirar proveito tendo em vista que algumas jogadas que levam a vitória que sejam óbvias olhando normalmente podem ser mais dificeis de ver de primeira caso o tabuleiro esteja rotacionado, além de que, pelo nosso treinamento ser mais ágil, teríamos uma vantagem em questão de desenvolvimento (Desenvolveríamos mais rápido). Sim, posições simetricamente equivalentes deveriam ter o mesmo valor*

Exercício 1.3 - *Greedy Play*: Suponha que o jogador de aprendizado por reforço fosse ganancioso, isto é, sempre jogasse o movimento que ele avaliasse como o melhor. Ele aprenderia a jogar melhor, ou pior, do que um jogador não ganancioso? Que problemas poderiam ocorrer?

Resolução: *Poderia ocorrer de que ele começasse a sempre utilizar uma mesma sequência de jogadas que foi a que maximizaram o seu ganho, de forma que ele se torne previsível, ou seja, aprenda pior*

Exercício 1.4 - *Learning from Exploration*: Suponha que as atualizações de aprendizado ocorressem após todos os movimentos, incluindo os movimentos exploratórios. Se o parâmetro de taxa de aprendizado for reduzido adequadamente ao longo do tempo (mas não a tendência de explorar), então os valores dos estados convergiriam para um conjunto diferente de probabilidades. Quais (conceitualmente) são os dois conjuntos de probabilidades computados quando aprendemos e quando não aprendemos a partir de movimentos exploratórios? Supondo que continuemos a fazer movimentos exploratórios, qual conjunto de probabilidades seria melhor aprender? Qual resultaria em mais vitórias?

Resolução: *Quando não aprendendemos com os movimentos de exploração, temos um conjunto* ***ótimo*** *de probabilidades, de forma que sempre vai levar o meu modelo ao resultado esperado. Faz sentido o melhor ser o conjunto ótimo, mesmo que possa levar a alguns casos indesejados, e eu também acredito que o que mais geraria vitórias seria esse*

Exercício 1.5 - *Other Improvements*: Você consegue pensar em outras maneiras de melhorar o jogador de aprendizado por reforço? Consegue pensar em alguma forma melhor de resolver o problema do jogo-da-velha, conforme foi proposto?

Resolução: *Eu penso em penalizar as jogadas que causam perca certa na próxima jogada, de forma que ele vai perceber que aquela jogada nunca é correta de se fazer, de tal modo que ele vai convergir para um aprendizado ideal mais rápido*