Inteligência Artificial - Lista de Exercícios 3

Prof. Fabrício Olivetti de França, Prof. Denis Fantinato

3º Quadrimestre de 2018

Inteligência Artificial

Exercícios

Baseados em:

Capítulo 16 e 17 (MDP): (16.2, 16.3, 16.6, 16.7, 16.8, 17.1, 17.3, 17.5, 17.8, 17.10)

Capítulo 21 (Aprendizado por Reforço): (21.1, 21.5, 21.6, 21.10)

Tradução

Capítulo 16:

16.2) Chris avalia 4 carros usados antes de comprar aquele que tem a maior utilidade esperada. Pat avalia 10 carros e também deseja comprar o de maior utilidade esperada. Considerando condições de análise iguais para Chris e Pat, quem tem mais chances de comprar o melhor carro? Quem tem mais chances de se arrepender da qualidade do carro? Você conseguiria quantificar essas chances em termos do desvio padrão da qualidade esperada?

16.3) Em 1713, Nicolas Bernoulli formulou um puzzle, agora conhecido como "o paradoxo de São Petersburgo", que funciona da seguinte maneira. Você tem a oportunidade de jogar um jogo no qual uma moeda justa é lançada repetidamente até que o resultado seja "cara". A primeira vez que aparecer "cara" no n-ésimo lançamento, você ganha 2^n reais.

- a. Mostre que o valor monetário esperado desse jogo é infinito.
- b. Quanto você pagaria para jogar esse jogo?
- c. Daniel Bernoulli, primo de Nicolas, resolveu o aparente paradoxo em 1738 sugerindo que a utilidade monetária fosse medida usando uma escala logarítmica (ou seja, $U(S_n) = a \log_2 n + b$, em que S_n é o estado com $R\$n$). Qual é a utilidade esperada para o jogo sob essa suposição.
- d. Qual seria a quantidade máxima que alguém pagaria para jogar o jogo, assumindo que essa pessoa já possui $R\$k$?

16.6) Prove que as preferências $B \succ A$ e $C \succ D$ no paradoxo de Allais (página 620) viola o axioma de substituição.

16.7) Considere o paradoxo de Allais descrito na página 620: um agente que prefere B em relação a A, e C em relação a D (assumindo o maior valor monetário esperado (EMV)) não está agindo racionalmente, de acordo com a teoria da utilidade. Você acha que isso seria um problema para o agente, um problema para a teoria, ou não há problema algum? Justifique.

16.8) Bilhetes para uma loteria custam $R\$1$. Existem dois possíveis prêmios: $R\$10{,}00$ com probabilidade $1/50$ e $R\$1.000.000{,}00$ com probabilidade $1/2.000.000$. Qual é o valor monetário esperado para o bilhete da loteria? Por qual valor (se houver) seria razoável comprar um bilhete? Seja preciso - mostre uma equação envolvendo utilidades. Você pode assumir que seu saldo atual é $R\$k$ e que $U(S_k) = 0$. Você também pode assumir que $U(S_{k+10}) = 10 \times U(S_{k+1})$, mas você não pode assumir nada sobre $U(S_{k+1.000.000})$. Estudos sociológicos mostram que pessoas com renda baixa compram um número de bilhetes de forma não proporcional. Você acha que isto é resultado de uma decisão ruim ou de um uso de uma função utilidade diferente?

Capítulo 17:

17.1) Para o mundo 4×3 mostrado na Figura 17.1, calcule quais quadrados podem ser alcançados a partir de $(1, 1)$, pela sequência de ações $[cima, cima, direita, direita, direita]$ e com quais probabilidades.

17.3) Suponha que nós definimos a utilidade de uma sequência de estados para ser a *máxima* recompensa obtida em qualque estado da sequência. Mostre que esta função utilidade não resulta em preferências estacionárias entre sequências de estado. Seria possível definir uma função utilidade em estados tal que a decisão por máxima utilidade esperada (MEU) resulta no comportamento ótimo do agente?

17.5) Para o ambiente mostrado na Figura 17.1, encontre os valores limites (*threshold*) para $R(s)$ tal que a política ótima muda quando o valor limite é cruzado. Você irá precisar calcular a política ótima e seu valor fixo $R(s)$. Dica: Prove que o valor de qualquer política fixa varia linearmente com $R(s)$.

17.8) Considere o mundo 3×3 mostrado na Figura 17.14(a). O modelo de transição é mesmo que aquele do mundo 4×3 na Figura 17.1: 80% do tempo o agente vai na direção que ele seleciona; no restante do tempo ele vai em direção ortogonal àquela pretendida.

Implemente o algoritmo de valor-iteração para este mundo para cada um dos valores de r abaixo. Use recompensas com desconto com um fator de $\gamma = 0.99$. Mostre a política obtida em cada caso. Explique intuitivamente por que o valor de r leva a cada política.

- a. $r = 100$
- b. $r = -3$
- c. $r = 0$

- d. $r = +3$

17.10) Considere um MDP com três estados, $(1, 2, 3)$, com recompensas $-1, -2, 0$, respectivamente. O estado 3 é um estado terminal. Nos estados 1 e 2, há duas ações possíveis: a e b. O modelo de transição é como segue:

- No estado 1, a ação a move o agente para o estado 2 com probabilidade 0.8 ou faz com que o agente permaneça no mesmo estado com probabilidade 0.2.
- No estado 2, a ação a move o agente para o estado 1 com probabilidade 0.8 ou faz com que o agente permaneça no mesmo estado com probabilidade 0.2.
- No estado 1 ou estado 2, a ação b move o agente para o estado 3 com probabilidade 0.1 ou faz com que o agente permaneça no mesmo estado com probabilidade 0.9.

Com base nisso, responda as seguintes questões:

- a. O que pode ser determinado *qualitativamente* sobre a política ótima nos estados 1 e 2?

- b. Aplique o algortimo de política-iteração, mostrando cada passo, para determinar a política ótima e os valores dos estados 1 e 2. Assuma que a política inicial tem ação b em ambos os estados.

- c. O que acontece com o algoritmo de política-iteração se a política inicial tem ação a em ambos os estados? Aplicar desconto ajuda? A política ótima depende do fator de desconto?

Capítulo 21:

21.1) Implemente um agente com aprendizado passivo em um ambiente simples, como no mundo 4×3. Para o caso com modelo de ambiente inicialmente desconhecido, compare o desempenho do aprendizado usando a estimativa direta de utilidade e a diferença temporal (TD). Faça a comparação para a política ótima e para várias políticas aleatórias. Para qual caso a estimativa de utilidade converge mais rápido? O que acontece se o tamanho do ambiente aumenta? Teste ambientes com e sem obstáculos.

21.5) Implemente um agente com aprendizado por reforço que usa a estimativa direta da utilidade. Faça duas versões: uma com a representação tabular e outra usando a função de aproximação (Equação 21.10 do livro). Compare o desempenho deles em três ambientes:

- a. No mundo 4×3.
- b. Em um mundo 10×10 sem obstáculos e com recompensa $+1$ em $(10, 10)$.
- c. Em um mundo 10×10 sem obstáculos e com recompensa $+1$ em $(5, 5)$.

21.6) Tente encontrar características adequadas ($f(s)$) para aprendizado por reforço em mundos tabulares/grade (generalizações do mundo 4×3) que contenham múltiplos obstáculos e múltiplos estados terminais com recompensas $+1$ ou -1.

21.10) Aprendizado por reforço é um modelo abstrato apropriado para evolução? Que conexão existe (se existir) entre sinais cerebrais de recompensa (p.ex.: felicidade, prazer, dor) e uma medida de desempenho evolucionária?