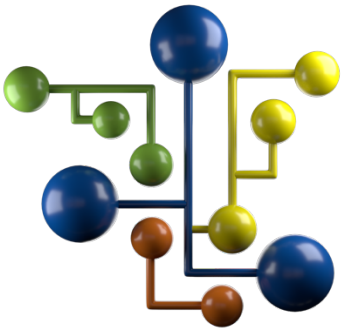


Data Science Academy

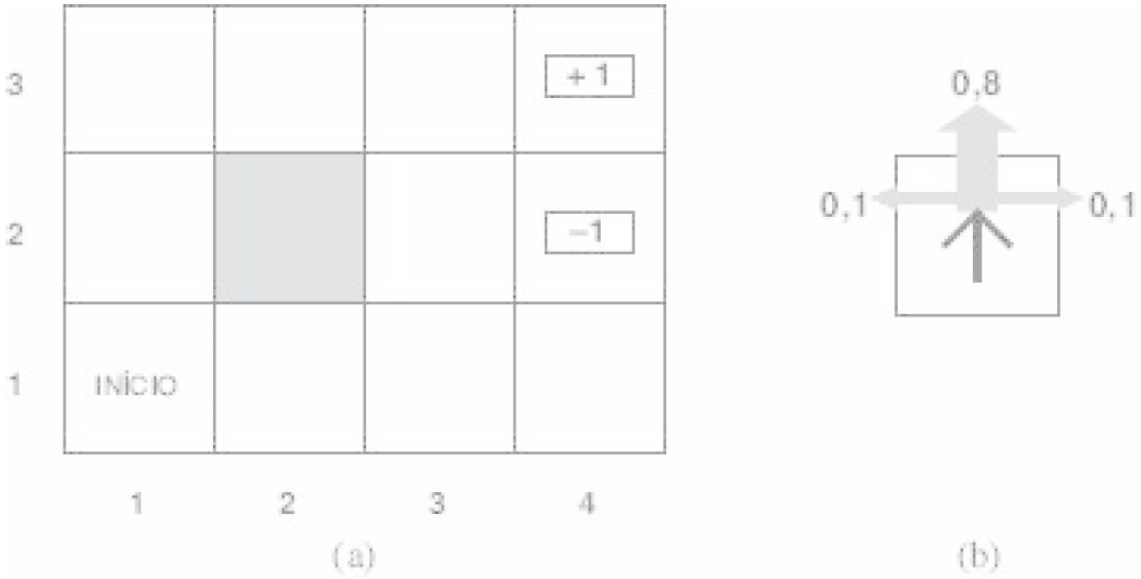
www.datascienceacademy.com.br

Introdução à Inteligência Artificial

Problemas de Decisão Sequencial

Abordaremos agora as questões computacionais envolvidas na tomada de decisões em ambiente estocástico. Enquanto anteriormente estávamos preocupados com problemas de decisão instantânea ou episódica, em que a utilidade do resultado de cada ação era bem conhecida, aqui vamos nos preocupar com problemas de decisão sequencial, em que a utilidade do agente depende de uma sequência de decisões. Problemas de decisão sequencial incorporam utilidades, incerteza e percepção, e incluem os problemas de busca e planejamento como casos especiais.

Suponha que um agente esteja situado no ambiente 4 × 3 mostrado na figura a abaixo. Começando no estado inicial, ele deve escolher uma ação em cada passo de tempo. A interação com o ambiente termina quando o agente alcança um dos estados objetivos, marcados com +1 ou –1. Assim como para problemas de busca, as ações disponíveis para o agente em cada estado são dadas por AÇÕES(s), algumas vezes abreviado como A(s); no ambiente 4 × 3, as ações em todos os estados são **Acima**, **Abaixo**, **Esquerda** e **Direita**. Vamos supor, por enquanto, que o ambiente seja completamente observável, de forma que o agente sempre saiba onde está.



Se o ambiente fosse determinístico, seria fácil encontrar uma solução: [Acima, Acima, Direita, Direita, Direita]. Infelizmente, o ambiente nem sempre responderá como esperado com essa solução porque as ações são pouco confiáveis. O modelo específico de movimento estocástico que adotamos está ilustrado na figura b. Cada ação alcança o efeito pretendido com probabilidade 0,8, mas, no restante do tempo, a ação move o agente em ângulos retos até a direção pretendida.

Além disso, se o agente bater em uma parede, ele permanecerá no mesmo quadrado. Por exemplo, a partir do quadrado inicial (1,1), a ação **Acima** move o agente para (1,2) com probabilidade 0,8, mas, com probabilidade 0,1, ele se move para a direita até (2,1) e, com

probabilidade 0,1, ele se move para a esquerda, choca-se com a parede e fica em (1,1). Em tal ambiente, a sequência [Acima, Acima, Direita, Direita, Direita] contorna a barreira e alcança o estado de meta em (4,3) com probabilidade 0,85 = 0,32768. Também existe uma pequena chance de atingir acidentalmente a meta indo por outro caminho, com probabilidade 0,14 × 0,8, dando um total geral igual a 0,32776 (veja também o Exercício 17.1).

O modelo de transição (ou apenas "modelo", quando não gerar confusão) descreve o resultado de cada ação em cada estado. Aqui, o resultado é estocástico, então escrevemos P(s' | s, a) para indicar a probabilidade de alcançar o estado s' se a ação a for feita no estado s.

Vamos supor que a probabilidade de alcançar s' a partir de s depende apenas de s, e não do histórico de estados anteriores. No momento, você pode pensar em P(s' | s, a) como uma grande tabela tridimensional contendo probabilidades. O modelo de transição pode ser representado como uma rede bayesiana dinâmica.

Para completar a definição do ambiente de tarefa, devemos especificar a função utilidade para o agente. Como o problema de decisão é sequencial, a função utilidade dependerá de uma sequência de estados — um histórico do ambiente —, em vez de depender de um único estado.

Vamos simplesmente estipular que, em cada estado s, o agente recebe uma recompensa R(s), que pode ser positiva ou negativa, mas deve ser limitada. Para nosso exemplo específico, a recompensa é –0,04 em todos os estados, exceto os estados terminais (que têm recompensas +1 e – 1). A utilidade de um histórico do ambiente é simplesmente (por enquanto) a soma das recompensas recebidas. Por exemplo, se o agente alcançar o estado +1 depois de 10 passos, sua utilidade total será 0,6. A recompensa negativa igual a –0,04 dá ao agente um incentivo para alcançar (4,3) depressa e, assim, nosso ambiente é uma generalização estocástica dos problemas de busca do Capítulo 2. Outro modo de dizer isso é afirmar que o agente não aprecia viver nesse ambiente e, portanto, quer deixá-lo assim que possível.

Para resumir: um problema de decisão sequencial para um ambiente completamente observável, estocástico, com um modelo de transição de Markov e recompensas aditivas, é chamado de processo de decisão de Markov ou **MDP (Makov Decison Process),** e consiste de um conjunto de estados (com estado inicial s0); um conjunto de AÇÕES(s) de ações aplicáveis em cada estado; um modelo de transição P(s' | s, a) e uma função de recompensa R(s).

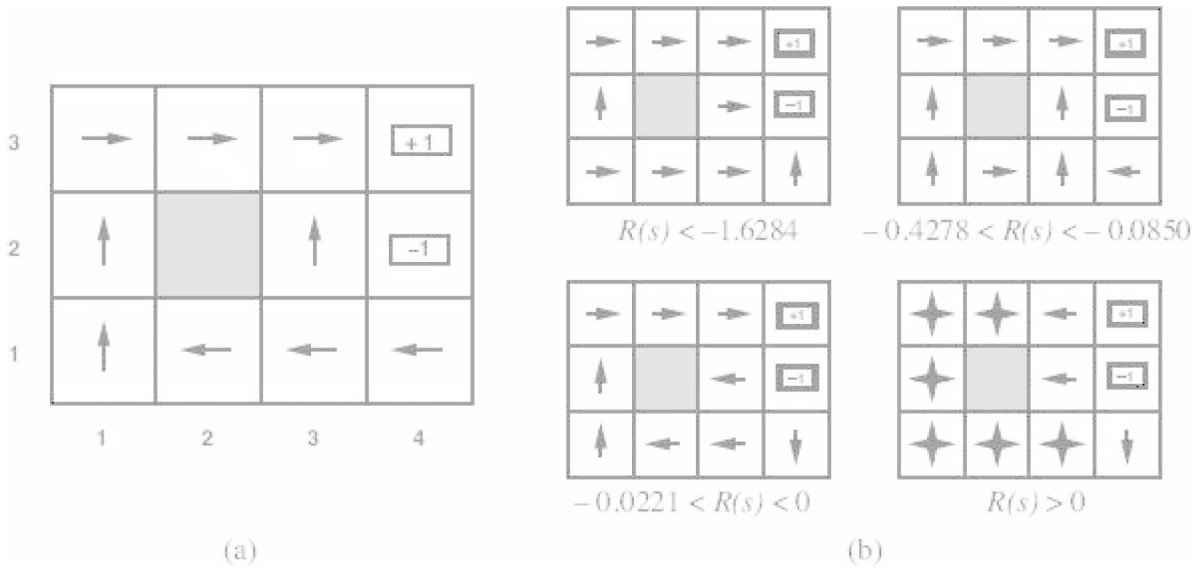
A próxima questão é definir qual seria a aparência de uma solução para o problema. Vimos que qualquer sequência fixa de ações não resolverá o problema porque o agente poderia acabar em um estado diferente da meta. Desta forma, uma solução tem de especificar o que o agente deve fazer para qualquer estado que o agente possa alcançar. Uma solução desse tipo é chamada de política. Normalmente, indicamos uma política por π e π(s) é a ação recomendada

pela política π para o estado s. Se o agente tiver uma política completa, não importará o resultado de qualquer ação, o agente sempre saberá o que fazer em seguida.

Toda vez que uma dada política for executada a partir do estado inicial, a natureza estocástica do ambiente poderá levar a um histórico de ambiente diferente. A qualidade de uma política é, portanto, medida pela utilidade esperada dos históricos de ambientes possíveis gerados por essa política.

Uma política ótima é uma política que produz a utilidade esperada mais alta. Usamos π* para indicar uma política ótima. Dado π*, o agente decide o que fazer consultando sua percepção atual, que informa o estado atual s, e depois executando a ação π*(s). Uma política representa explicitamente a função do agente e, portanto, é uma descrição de um agente reflexivo simples, calculada a partir das informações usadas por um agente baseado na utilidade.

Uma política ótima é mostrada na figura a abaixo. Observe que, como o custo de dar um passo é bastante pequeno em comparação com a penalidade por terminar em (4,2) por acidente, a política ótima para o estado (3,1) é conservadora. A política recomenda seguir o caminho longo, em vez de tomar o atalho e se arriscar a entrar em (4,2).



O equilíbrio entre risco e recompensa muda dependendo do valor de R(s) para os estados não terminais. A figura b mostra políticas ótimas para quatro intervalos diferentes de R(s). Quando R(s) ≤ –1,6284, a vida é tão difícil que o agente vai direto para a saída mais próxima, ainda que a saída tenha o valor –1. Quando –0,4278 ≤ R(s) ≤ –0,0850, a vida é bastante desagradável; o agente toma a rota mais curta até o estado +1 e está disposto a correr o risco de cair no estado –1 por acidente. Em particular, o agente toma o atalho a partir de (3,1). Quando a vida é apenas ligeiramente ruim (–0,0221 < R(s) < 0), a política ótima não assume absolutamente nenhum risco. Em (4,1) e (3,2) o agente segue diretamente para fora do

estado –1, de forma que não possa cair nesse estado por acidente, embora isso signifique bater a cabeça contra a parede várias vezes. Finalmente, se R(s) > 0, a vida positivamente é agradável e o agente evita ambas as saídas. Desde que as ações em (4,1), (3,2) e (3,3) sejam as que estão representadas, toda política é ótima, e o agente obtém recompensa total infinita porque nunca entra em estado terminal. Surpreendentemente, verificamos que existem seis outras políticas ótimas para vários intervalos de R(s).

O equilíbrio cuidadoso entre risco e recompensa é uma característica dos MDPs que não surge em problemas de busca determinística; além disso, é uma característica de muitos problemas de decisão do mundo real. Por essa razão, os MDPs foram estudados em vários campos, inclusive em IA, pesquisa operacional, economia e teoria de controle. Foram propostas dezenas de algoritmos para calcular políticas ótimas.

Referências:

Livro: Inteligência Artificial
Autor: Peter Norvig