Capítulo 1

Tomada de Decisão Sequencial

1.1 Processo Markoviano de Decisão

Um processo markoviano de decisão (*Markovian Decision Process* – MDP) é definido por uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$ onde:

- $s \in \mathcal{S}$ são estados possíveis;
- $a \in \mathcal{A}$ são ações possíveis;
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ é a função de transição; e
- $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ é a função recompensa.

Um MDP define o seguinte processo para todo tempo $t \geq 0$:

1. o agente percebe o estado s_t;
2. o agente escolhe e executa uma ação a_t;
3. o ambiente transita para o próximo estado s_{t+1} com probabilidade $\Pr(s_{t+1} = s' | a_t = a, s_t = s) = T(s, a, s')$; e

4. o agente recebe uma recompensa $R(s_t, a_t, s_{t+1})$.

Um agente que escolhe as ações a_t deve fazer isso de modo a buscar recompensas positivas e evitar recompensas negativas.

Estado Inicial e Horizontes

Na seção anterior foi definido um processo, mas nada foi falado sobre como esse processo se inicializa e nem quando o mesmo termina.

Com relação ao estado inicial, pode-se defini-lo como um estado inicial único, simplesmente denominado s_0, nesse caso, um MDP seria uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R, s_0 \in \mathcal{S} \rangle$. Outra forma de definir o estado é considerar uma distribuição de probabilidade para o estado inicial descrita por uma função de estado inicial $I : \mathcal{S} \to [0, 1]$, tal que $\Pr(s_0 = s) = I(s)$, nesse caso, um MDP seria uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R, I \rangle$.

Usualmente, a área de planejamento considera um estado inicial s_0 e faz uso dessa informação para obter soluções com menor complexidade computacional e menor uso de memória. Já, no Aprendizado por Reforço, é mais comum não considerar estados inicias, mas, como as experiências vem de experimentos reais, esses experimentos vêm de alguma distribuição inicial e influencia o resultado obtido.

Com relação ao horizonte para o qual se estende o processo, pode-se ter três opções: horizonte finito, horizonte infinito, e horizonte indeterminado. No caso do horizonte finito, é definido um horizonte máximo N, de tal forma que o processo continua enquanto $t \leq N$. No caso do horizonte infinito, o processo nunca para. Finalmente, no horizonte indeterminado, considera-se estados absorvedores, usualmente um conjunto de estados metas $\mathcal{G}$, de tal forma que o processo acaba quando $s_t \in \mathcal{G}$, nesse caso, um MDP seria uma tupla $\langle \mathcal{S}, \mathcal{A}, T, R, s_0 \in \mathcal{S}, \mathcal{G} \rangle$. O caso de horizonte indeterminado também é o mais comum na área de planejamento.

Um estado absorvedor é um estado que nunca sai dele e produz recompensa nula, isto é, para todo $s \in \mathcal{G}$ e $s' \neq s \in \mathcal{S}$, tem-se que $T(s, a, s) = 1$ e $T(s', a, s) = 0$ e $R(s, a, s) = R(s, a, s') = 0$.

Recompensas versus Custos

Como já foi falado antes, a área de planejamento usualmente considera um estado inicial e estados metas. Outra variação tipicamente utilizada na área de planejamento é a ideia de custo. O custo seria as recompensas negativas.

Quando as recompensas envolvidas são todas negativas e considera-se estados metas, tem-se o problema de Caminho Estocástico mais Curto (*Shortest Stochastic Path* – SSP). Nesse caso, a função recompensa é trocada por uma função custo $C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$, assim, um SSP é definido por $\langle \mathcal{S}, \mathcal{A}, T, C, s_0 \in \mathcal{S}, \mathcal{G} \rangle$.

Outras variações bastante encontradas nos trabalhos é com relação ao argumento da função recompensa. Nos trabalhos, é bastante comum encontrar as seguintes variações: $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ e $R : \mathcal{S} \to \mathbb{R}$.

Finalmente, pode-se considerar uma recompensa estocástica, na qual a recompensa r_t no tempo t vem de uma distribuição de probabilidade $\Pr(r_t = x | s_t = s, a_t = a, s_{t+1} = s')$.

Ações versus Estados

Outra variação bastante comum em trabalhos de Planejamento ou Aprendizado por Reforço é restringir as ações disponíveis para cada estado, isto é, considera-se uma função $A : \mathcal{S} \to 2^{\mathcal{A}}$ e as ações escolhidas no tempo t são limitadas a $a_t \in A(s_t)$.

Observação Parcial

Uma formalização mais completa que o MDP e também mais realista para vários problemas é considerar observações parciais. No

processo descrito para o MDP, o agente observa diretamente o estado s_t, no caso da formalização com observações parciais, o agente observa o estado indiretamente por uma função de observação.

Dado um conjunto de observações $\mathcal{O}$, em cada tempo t o agente faz uma observação $o_t \in \mathcal{O}$. A relação da observação o_t com o estado s_t é dada por uma função de observação $O : \mathcal{S} \to (\mathcal{O} \to [0, 1])$ ou $O : \mathcal{S} \times \mathcal{O} \to [0, 1]$, tal que $\Pr(o_t = o|s_t) = O(s, o)$.

1.2 Políticas

As ações a_t podem ser escolhidas de forma arbitrária e são definidas por uma política de ações π, ou, simplesmente, política. Essencialmente, uma política mapeia situações em uma ação.

A execução de uma política associada ao processo regido por uma função de transição T, gera histórias $h_t = s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t$.

Políticas podem ser classificadas sob três aspectos:

- estacionária ou não-estacionária,
- markoviana ou não-markoviana; e
- determinista ou probabilística.

Uma política estacionária é igual para todo tempo $t \geq 0$, enquanto uma política não-estacionária vária com o tempo. Uma política markoviana depende apenas do estado atual, isto é, escolhe a ação a_t com base apenas no estado atual s_t, enquanto uma política não-markoviana escolhe a ação a_t com base no histórico $s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t$. Finalmente, uma política determinista escolhe sempre a mesma ação quando em uma mesma situação, enquanto uma política probabilística escolhe uma ação segundo uma distribuição de probabilidade.

A classe de políticas mais simples é classe de políticas estacionárias, markovianas e deterministas e simplesmente mapeiam

estados em ações, isto é, $\pi : \mathcal{S} \to \mathcal{A}$. Uma política não-estacionária, markoviana e determinista mapeia pares estados-tempo em ações, isto é, $\pi : \mathcal{S} \times \mathbb{N} \to \mathcal{A}$. Uma política estacionária, não-markoviana e determinista mapeia história em ações, isto é, $\pi : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^* \times \mathcal{S} \to \mathcal{A}$. Uma política estacionária, markoviana e probabilística mapeia estados em distribuições de probabilidade sobre ações, isto é, $\pi : \mathcal{S} \to (\mathcal{A} \to [0, 1])$ ou $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$.

1.3 Avaliando Políticas

Ao executar uma política π um agente escolhe ações $a_t = \pi(s_t)$ e gera um histórico $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$, que pode ser finito ou infinito, dependendo do tipo de horizonte considerado. Há três formas básicas na literatura para avaliar uma política: média, somatória e desconto.

Se considerarmos todos os estamos como potenciais estados iniciais, a avaliação de uma política consiste em criar uma função valor $V^\pi : \mathcal{S} \to \mathbb{R}$ que especifica um escalar para cada estado. Dessa forma, dado um estado inicial s_0 arbitrário, pode-se comparar duas políticas π e π' comparando os valores $V^\pi(s_0)$ e $V^{\pi'}(s_0)$.

Definição 1. *O valor de uma política segundo o critério de recompensa média é dado por:*

$$V^\pi(s) = \lim_{N \to \infty} \frac{1}{N} E\left[\sum_{t=0}^{N-1} r_t \middle| s_0 = s, \pi \right].$$

A recompensa média pode ser utilizada apenas para horizontes infinitos, no qual o processo nunca acaba. Note que o horizonte infinito é especificado pelo limite $N \to \infty$. Ainda, note que se as recompensas se extinguirem, como no caso de horizonte indeterminado, o limite tenderá a zero para qualquer política.

Definição 2. *O valor de uma política segundo o critério de recompensa acumulada esperada é dado por:*

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} r_t \middle| s_0 = s, \pi\right].$$

A recompensa acumulada esperada pode ser utilizada apenas para horizontes finitos ou indeterminados. Como todas as recompensas são acumuladas, esse valor será finito apenas se as recompensas cessarem. Dessa forma, esse critério não pode ser utilizado para horizontes infinitos.

Definição 3. *O valor de uma política segundo o critério de recompensa acumulada descontada é dado por:*

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, \pi\right],$$

onde $\gamma < 1$ é um fator de desconto.

A recompensa acumulada descontada garante que, quando a recompensa é limitada, o acumulo descontado seja sempre finito, independente do tipo de horizonte. Além disso, quando $\gamma \to 1$, esse critério se aproxima do critério de recompensa média se o horizonte for infinito ou se aproxima do critério de recompensa acumulada se o horizonte for finito ou indeterminado. Portanto, utilizar fator de desconto próximo a 1, pode ser uma boa aproximação para qualquer dos três horizontes considerados.

Além de ser um truque matemático para garantir convergência da recompensa acumulada, o fator de desconto γ também pode ser interpretado semanticamente. Primeiro, pode-se pensar o fator de desconto como uma importância para recompensas mais próximas no tempo, sendo que quanto menor o fator de desconto, menos vale as recompensas obtidas em futuro longe. Segundo, pode-se pensar o fator de desconto como uma inflação, que diminui o valor da

recompensa no tempo. Terceiro, pode-se pensar o fator de desconto como uma chance de processo continuar.