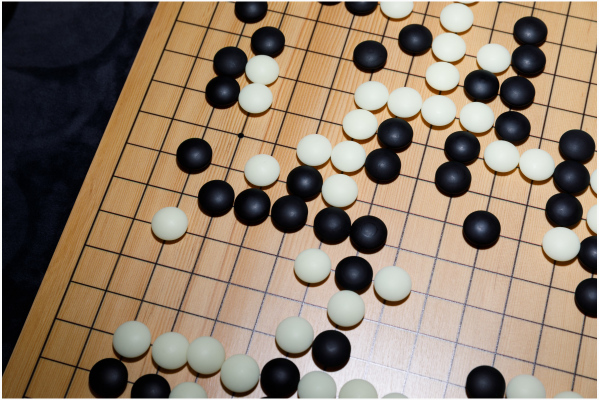
ACH2016 - Inteligência Artificial
Aula 13 - Processos Markovianos de Decisão

Valdinei Freire da Silva

valdinei.freire@usp.br - Bloco A1 100-O

Russell e Norvig, Capítulo 17

AlphaGo



- Solução: busca aleatória heurística + aprendizado por reforço + redes neurais
- Documentário: AlphaGo

Resolução de Problemas

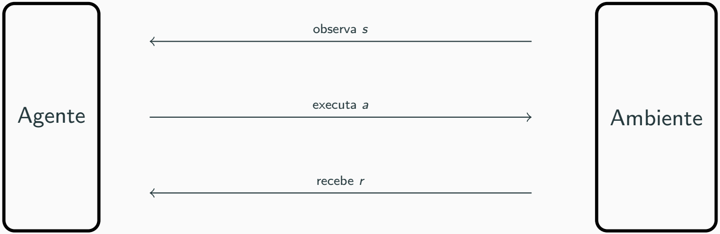
Ambientes: completamente observável, único agente, conhecido, determinístico, discreto, sequencial, estático.

Formulação de Problemas

- estado inicial: é o estado no tempo $t = 0$.
- ações possíveis: a função $\mathrm{ACTIONS}(s)$ retorna o conjunto de ações que podem ser executadas no estado s.
- modelo de transição: a função $\mathrm{RESULT}(s, a)$ retorna o estado resultante de aplicar a ação a no estado s.
- teste de meta: a função $\mathrm{GOAL}(s)$ determina se o estado s é um estado meta.
- solução: plano ρ que consiste em uma sequência de ações $\rho = a_0, a_1, \ldots, a_{T-2}, a_{T-1}$.

Ambientes Probabilístico

Ambientes: completamente observável, único agente, conhecido, **probabilístico**, discreto, sequencial, estático.



Propriedade de Markov: O próximo estado no processo depende apenas do estado atual e da ação escolhida.

Processo Markoviano de Decisão

Markov Decision Process (MDP) é definido pela tupla $\langle \mathcal{S}, \mathcal{A}, T, R \rangle$

- $\mathcal{S}$ é o conjunto de estados
- $\mathcal{A}$ é o conjunto de ações
- $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ é a função de transição
- $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ é a função recompensa

Processo: em cada tempo t

- o processo encontra-se no estado s_t
- uma ação a_t é escolhida
- a recompensa $r_t = r(s_t, a_t)$ é gerada
- o processo transita para um estado s' com distribuição $\Pr(s_{t+1} = s' | s_t = s, a_t = a) = T(s, a, s')$

$$\Pr(s_{t+1} | s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = \Pr(s_{t+1} | s_t, a_t)$$

Objetivo

Objetivo: buscar recompensas positivas e evitar recompensas negativas.

Solução: política de ação que escolhe ações em cada situação.

Política Estacionária: qual ação executar no estado s

$$\pi : \mathcal{S} \rightarrow \mathcal{A}$$

Política Não-estacionária: qual ação executar em s, no tempo t

$$\pi : \mathcal{S} \times \mathbb{N} \rightarrow \mathcal{A}$$

Política Probabilística: qual a probabilidade de executar a em s

$$\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$$

MDPs e Horizonte

Horizonte Finito: o processo termina em um tempo τ definido *a priori*, dessa forma um MDP é definido por $\langle \mathcal{S}, \mathcal{A}, T, R, \tau \rangle$.

Horizonte Indeterminado: considera-se a existência de um conjunto de estados terminais $\mathcal{T} \subset \mathcal{S}$ e se $s \in \mathcal{T}$ então s é um estado absorvente, isto é,

$$\forall s \in \mathcal{T}, a \in \mathcal{A} \, \Pr(s|s,a) = 1 \text{ e } r(s,a) = 0.$$

Dessa forma, pode-se definir uma variável aleatória τ^{π} que indica quando o processo acaba e que depende da política executada.

Horizonte Infinito: o processo nunca termina, podendo gerar recompensas diferentes de zero em qualquer momento.

Medida de Desempenho

Na interação do agente com o ambiente ocorrem uma sequência de recompensas $r_0, r_1, \ldots$ que são acumuladas na variável aleatória R

Horizonte Finito: somatório de recompensas até o fim do processo no tempo τ, isto é,

$$R = \sum_{t=0}^{\tau-1} r_t$$

Horizonte Indeterminado: somatório de recompensas até encontrar um estado absorvente no tempo τ, isto é,

$$R = \sum_{t=0}^{\tau-1} r_t$$

Note que pode acontecer de o agente não chegar em um estado absorvente e o processo não acabar. Nesse caso, a política executada é imprópria (*non-proper*)

Medida de Desempenho

Horizonte Infinito: como o processo nunca termina, deve-se adotar algum critério para que o somatório de recompensas não divirja.

Somatório descontado: $R = \lim_{M \to \infty} \sum_{t=0}^{M} \gamma^t r_t$, para fator de desconto $\gamma \in (0, 1)$

Recompensa média: $R = \lim_{M \to \infty} \frac{1}{M} \sum_{t=0}^{M-1} r_t$

Fator de Desconto - Significados

- as recompensas recebidas mais distantes no tempo importam menos
- se for valor monetário, pode-se pensar em uma inflação
- a cada tempo existe uma chance γ de o agente continuar vivo

Fator de Desconto - Vantagens

Vantagens:

- pode ser utilizado tanto em ambientes de horizonte Indeterminado, como em ambientes de horizonte infinito
- quando $\gamma \rightarrow 1$, a solução converge para a solução de recompensa média em ambientes de horizonte infinito
- quando $\gamma \rightarrow 1$, a solução converge para a solução sem desconto em ambientes de horizonte indeterminado

Desvantagens:

- para qualquer valor de γ escolhido, pode-se construir um MDP tal que a solução descontada não é própria (*proper*)
- a solução sem desconto é mais genérica que a solução com desconto, isto é, todo algoritmo que encontra solução para desempenho sem desconto pode ser utilizado para resolver problemas considerando desempenho com desconto

Valor de Política

Considere que um agente ao executar uma política π obtém recompensas $r_0, r_1, \ldots$, então pode-se considerar a seguinte variável aleatória

$$R^{\pi,\gamma} = \lim_{M \to \infty} \sum_{t=0}^{M-1} \gamma^t r_t = \sum_{t=0}^{\infty} \gamma^t r_t$$

O valor da política π pode ser definido por:

$$V^{\pi,\gamma} = \mathrm{E}[R^{\pi,\gamma}]$$

Função Valor

Definição O valor para todo $s \in \mathcal{S}$ ao seguir π é dado por[1]:

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

Se π for estacionária e determinista, pode-se encontrar os valores resolvendo um sistema de equação linear:

$$\begin{aligned} V^{\pi}(s) &= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s', \pi\right] \\ &= r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V^{\pi}(s') \end{aligned}$$

[1] Sempre que possível o fator γ será deixado subentendido.

Função Valor - Notação Vetorial

Definição Considere as representações vetoriais das funções valor, recompensa e transição para uma política π, onde:

$$\mathrm{V}^\pi = \begin{bmatrix} V^\pi(1) \\ V^\pi(2) \\ \vdots \\ V^\pi(|\mathcal{S}|) \end{bmatrix}, \mathrm{r}^\pi = \begin{bmatrix} r(1, \pi(1)) \\ r(2, \pi(2)) \\ \vdots \\ r(|\mathcal{S}|, \pi(|\mathcal{S}|)) \end{bmatrix}$$

$$\mathrm{T}^\pi = \begin{bmatrix} T(1, \pi(1), 1) & T(1, \pi(1), 2) & \cdots & T(1, \pi(1), |\mathcal{S}|) \\ T(2, \pi(2), 1) & T(2, \pi(2), 2) & \cdots & T(2, \pi(2), |\mathcal{S}|) \\ \vdots & \vdots & \ddots & \vdots \\ T(|\mathcal{S}|, \pi(|\mathcal{S}|), 1) & T(|\mathcal{S}|, \pi(|\mathcal{S}|), 2) & \cdots & T(|\mathcal{S}|, \pi(|\mathcal{S}|), |\mathcal{S}|) \end{bmatrix}$$

Sistema de equação linear em notação vetorial:

$$\mathrm{V}^\pi = \mathrm{r}^\pi + \gamma \mathrm{T}^\pi \mathrm{V}^\pi$$

Função Valor - Contração

Um operador $J : \mathcal{C} \to \mathcal{C}$ é uma contração se existe $A < 1$ tal que:

$$\left\| JV - JV^* \right\|_\infty \leq A \left\| V - V^* \right\|_\infty ,$$

onde $\|X\|_\infty = \sup_i X(i)$.

Escolha V_0 arbitrariamente e realize as seguintes atualizações:

$$V_i \leftarrow JV_{i-1},$$

então

$$\lim_{i \to \infty} V_i = V^*.$$

Avaliação de Política

Considere o seguinte operador:

$$J^{\pi}V = r^{\pi} + \gamma T^{\pi} V,$$

então J^{π} é uma contração, isto é:

$$\begin{aligned} \|J^{\pi}V - J^{\pi}V^{\pi}\|_{\infty} &= \|(r^{\pi} + \gamma T^{\pi} V) - (r^{\pi} + \gamma T^{\pi} V^{\pi})\|_{\infty} \\ &= \|\gamma T^{\pi} (V - V^{\pi})\|_{\infty} \leq \gamma \|V - V^{\pi}\|_{\infty} \end{aligned}$$

Teorema: convergência Seja V_0 arbitrária e defina $V_{n+1} = J^{\pi}V_n$, então:

$$\lim_{n\to\infty} V_n = V^{\pi}$$

Teorema: taxa de convergência

$$\|V^{\pi} - J^{\pi}V_n\|_{\infty} \leq \frac{\gamma}{1-\gamma} \|J^{\pi}V_n - V_n\|_{\infty}$$

Avaliação de Política

função Evaluate(MDP, V, π, ϵ)

repita

$\Delta \leftarrow 0$

para cada $s \in \mathcal{S}$

$v \leftarrow V(s)$

$$V(s) \leftarrow r(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V(s')$$

$\Delta \leftarrow \max\{\Delta, |v - V(s)|\}$

até $\Delta < \frac{1-\gamma}{\gamma} \epsilon$

retorna função valor V que aproxima V^{π} com erro máximo ϵ

Princípio da Otimalidade de Bellman

"Para um dado estado do sistema, a política ótima para os estados remanescentes é independente da política de decisão adotada em estados anteriores"

Programação Dinâmica:

- O problema pode ser dividido em etapas.
- Em cada etapa, é possível definir o estado da solução.
- A cada etapa, toma-se uma decisão que influencia o estado da etapa seguinte.

Programação Dinâmica - Exemplo

Considere uma mochila que pode carregar no máximo c quilos. Considere n itens que podem ser levados na mochila com peso $p_1, p_2, \ldots, p_n$, e valor $v_1, v_2, \ldots, v_n$. Encontre o conjunto de itens que caiba na mochila e que some o maior valor possível.
Todos as variáveis envolvidas no problema são números naturais.

Considere $V(j, w)$ o valor acumulado do subproblema considerando uma mochila que suporta w quilos e considera apenas os itens $1, 2, \ldots, j$.

$$V(j, w) = \max\{V(j-1, w), v_j + V(j-1, w-p_j)\}$$

Programação Dinâmica - MDP com Horizonte Finito

Considere $V^*(s, i)$ o valor esperado de iniciar o processo no estado s, tendo ainda i passos para agir e executando a política ótima, isto é

$$V^*(s, i) = \mathrm{E}\left[\sum_{t=0}^{i} r_t \middle| s_0 = s \right]$$

Tem-se que:

$$V^*(s, 0) = 0$$

$$V^*(s, i) = \max_{a \in \mathcal{A}} \left\{ r(s, a) + \sum_{s' \in \mathcal{S}} T(s, a, s') V(s', i - 1) \right\}$$

Função Valor estado-ação

Seja $Q^\pi(s,a)$ o valor esperado de iniciar o processo no estado s, executar a ação a e seguir com a política π dali em diante, isto é,

$$Q^\pi(s,a) = \mathrm{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t \,\middle|\, s_0 = s, a_0 = a, \pi\right]$$

Seja $Q^* = Q^{\pi^*}$ onde π^* é a política ótima. Então:

$$V^\pi(s) = Q^\pi(s,\pi(s))$$

$$V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a)$$

$$\pi^*(s) = \arg\max_{a\in\mathcal{A}} Q^*(s,a)$$

Backup de Bellman

Equação de Bellman para função V^* (sistema de equações não-lineares)

$$V^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V^*(s') \right\}$$

Equação de Bellman para função Q^* (sistema de equações não-lineares)

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') \max_{a' \in \mathcal{A}} Q^*(s',a')$$

$$\mathrm{Q}^*_a = \mathrm{r}^a + \gamma \mathrm{T}^a \max_{a' \in \mathcal{A}} \mathrm{Q}^*_{a'}$$

Backup de Bellman - Contração

O Backup de Bellman tem a propriedade de contração:

$$\left\|\left(\mathrm{r}^a + \gamma \mathrm{T}^a \max_{a' \in \mathcal{A}} \mathrm{Q}_{a'}\right) - \left(\mathrm{r}^a + \gamma \mathrm{T}^a \max_{a' \in \mathcal{A}} \mathrm{Q}^*_{a'}\right)\right\|_\infty \leq \gamma \left\|\mathrm{Q} - \mathrm{Q}^*\right\|_\infty$$

Defina o operador J como:

$$(JV)(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V(s') \right\}$$

Teorema: convergência Seja V_0 arbitrária e defina $\mathrm{V}_{n+1} = J\mathrm{V}_n$, então:

$$\lim_{n \to \infty} \mathrm{V}_n = \mathrm{V}^*$$

Teorema: taxa de convergência

$$\left\|\mathrm{V}^* - J\mathrm{V}_n\right\|_\infty \leq \frac{\gamma}{1-\gamma} \left\|J\mathrm{V}_n - \mathrm{V}_n\right\|_\infty$$

Algoritmo: Iteração de Valor

função ValueIteration(MDP,Q,ϵ)

repita

$\Delta \leftarrow 0$

para cada $s \in \mathcal{S}$

$V(s) \leftarrow \max_{a \in \mathcal{A}} Q(s,a)$

para cada $s \in \mathcal{S}$

$v \leftarrow V(s)$

para cada $a \in \mathcal{A}$

$$Q(s,a) \leftarrow r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s')V(s')$$

$\Delta \leftarrow \max\{\Delta, |v - \max_{a \in \mathcal{A}} Q(s,a)|\}$

até $\Delta < \frac{1-\gamma}{\gamma}\epsilon$

retorna função valor V que aproxima V^* com erro máximo ϵ

Algoritmo: Iteração de Política

função PolicyIteration(MDP,π,V,ϵ)

repita

$V \leftarrow$Evaluate(MDP,V,π,ϵ)

para cada $s \in \mathcal{S}$

$b(s) \leftarrow \pi(s)$

$$\pi(s) \leftarrow \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s')V(s') \right\}$$

até $\pi(s) = b(s) \forall s \in \mathcal{S}$

retorna política π que aproxima π^* com erro máximo ϵ