Algoritmos para MDP
Value Iteration and Policy Iteration

Valdinei Freire

(EACH - USP)

Exemplos

- Ações cardinais (N,S,L,O). Na linha "Det." as ações são deterministas. Na linha "Prob." as ações resultam com 0.5 de chance, caso contrário o agente fica parado. Custo de 1 por ação.

Prob.	s_0				G
Det.					

- Ações cardinais (N,S,L,O). As ações são deterministas, a menos do rio. No rio, as ações resultam com 0.5 de chance, caso contrário o agente volta para o estado inicial s_0.

s_0	rio	rio	rio	G

Programação Dinâmica

Uma política ótima tem a propriedade que seja qual for o estado inicial e decisões iniciais, as decisões restantes devem constituir uma política ótima com relação ao estado resultante da primeira decisão.

Problema da Mochila

- temos um número limitado de itens para colocar em uma mochila levando em consideração o peso de cada item e a capacidade do recipiente, e queremos maximizar o valor total de utilidade, sendo que a cada item é associado o seu valor de utilidade.
  - capacidade do recipiente: C
  - peso de cada item: $p_1, p_2, \ldots, p_3$
  - valor de utilidade de cada item: $v_1, v_2, \ldots, v_3$

Horizonte Finito

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

Política ótima é não-estacionária, determinista e markoviana

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

Função valor ótima $V^* : \mathcal{S} \times \mathbb{N} \rightarrow \mathbb{R}$, na qual $V^*(s, n)$ define a recompensada acumulada a partir do estado s e do passo $n < N$ seguindo a política ótima

Iteração de Valor: Horizonte Finito

1. defina $V(s, N) = 0$
2. faça para todo passo $n = N - 1$ to 0
   (a) para todo $s \in \mathcal{S}$

$$\begin{aligned} V(s,n) &= \max_{a \in \mathcal{A}} \left\{ \sum_{s' \in \mathcal{S}} T(s,a,s') \left[R(s,a) + V^*(s', n+1) \right] \right\} \\ &= \max_{a \in \mathcal{A}} \left\{ \overline{R}(s,a) + \sum_{s' \in \mathcal{S}} T(s,a,s') V^*(s', n+1) \right\} \end{aligned}$$

   (b) para todo $s \in \mathcal{S}$

$$\pi(s,n) = \arg\max_{a \in \mathcal{A}} \left\{ \overline{R}(s,a) + \sum_{s' \in \mathcal{S}} T(s,a,s') V^*(s', n+1) \right\}$$

3. retorne π

Horizonte Infinito

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

Política ótima é estacionária, determinista e markoviana

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

Função valor ótima $V^* : \mathcal{S} \rightarrow \mathbb{R}$, na qual $V^*(s)$ define a recompensada acumulada descontada a partir do estado s seguindo a política ótima

Operador de Bellman

Equação de Bellman

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

Definition 1. Para todo $s \in \mathcal{S}$, define-se o operador $\mathcal{T} : \Re^{|\mathcal{S}|} \to \Re^{|\mathcal{S}|}$ como:

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

Contração

Theorem 1. Um operador $\mathcal{F}$ é uma contração se existe $\beta < 1$ tal que $\|\mathcal{F}V' - \mathcal{F}V''\|_\infty \leqslant \beta\|V' - V''\|_\infty$. Em caso do operador $\mathcal{F}$ ser uma contração, então existe um único ponto fixo V^*, isto é, $\mathcal{F}V^* = V^*$ e $\lim_{n\to\infty} \mathcal{F}^n V = V^*$ para qualquer V.

Theorem 2. O operador de Bellman:

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

é uma contração.

Iteração de Valor: Horizonto Infinito

1. inicializa $V_0(s)$ arbitrariamente
2. faça para toda iteração $k \geqslant 0$
   (a) para todo $s \in \mathcal{S}$

$$V_{k+1}(s) = \max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V_k(s') \right\}$$

   enquanto $\|V_k - V_{k+1}\|_\infty < \epsilon$
3. retorne a política

$$\pi(s) = \arg\max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V_k(s') \right\}$$

Iteração de Valor: Horizonto Infinito

Theorem 3. Considere a função valore V_{k+1} retornada pelo algoritmo Iteração de Valor, então:

$$\|V_{k+1} - V^*\|_\infty \leqslant \frac{2\epsilon\gamma}{1-\gamma}.$$

Exemplos: $\gamma = 1$

- Ações cardinais (N,S,L,O). Na linha “Det.” as ações são deterministas. Na linha “Prob.” as ações resultam com 0.5 de chance, caso contrário o agente fica parado. Custo de 1 por ação.

Prob.	7	6	4	2	0
Det.	5	4	3	2	1

- Ações cardinais (N,S,L,O). As ações são deterministas, a menos do rio. No rio, as ações resultam com 0.5 de chance, caso contrário o agente volta para o estado inicial s_0.

5	4	3	2	1
6	6	5,5	4	0

Exemplos: $\gamma = 0.9$

- Ações cardinais (N,S,L,O). Na linha "Det." as ações são deterministas. Na linha "Prob." as ações resultam com 0.5 de chance, caso contrário o agente fica parado. Custo de 1 por ação.

Prob.	5.1687	4.5229	3.3058	1.8182	0
Det.	4.0951	3.439	2.71	1.9	1

- Ações cardinais (N,S,L,O). As ações são deterministas, a menos do rio. No rio, as ações resultam com 0.5 de chance, caso contrário o agente volta para o estado inicial s_0.

4.0951	3.439	2.71	1.9	1
4.6856	4.6561	4.3280	3.1085	0

Exemplos: $\gamma = 0.9$, custo nulo, $V_G = 1$

- Ações cardinais (N,S,L,O). Na linha “Det.” as ações são deterministas. Na linha “Prob.” as ações resultam com 0.5 de chance, caso contrário o agente fica parado. Valor terminal da meta $V_G = 1$.

Prob.	0.4831	0.5477	0.6694	0.8182	1
Det.	0.5905	0.6561	0.729	0.81	0.9

- Ações cardinais (N,S,L,O). As ações são deterministas, a menos do rio. No rio, as ações resultam com 0.5 de chance, caso contrário o agente volta para o estado inicial s_0.

0.5905	0.6561	0.729	0.81	0.9
0.5314	0.5344	0.5672	0.6891	1

Iteração de Política

Ideia:

- Avaliação de Política (*Policy Evaluation*)

- Melhoria de Política (*Policy Improvement*)

Critério de Convergência: Política atual não pode ser melhorada

Theorem 4. O algoritmo de Iteração de Política converge para a política ótima.

Avaliação de Política

Sistema de Equações Lineares:

$$V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V^\pi(s').$$

Notação Vetorial:

$$\begin{aligned} \mathbf{V}^\pi &= \mathbf{R}^\pi + \gamma \mathbf{T}^\pi \mathbf{V}^\pi \\ \mathbf{V}^\pi &= (\mathbf{I} - \gamma \mathbf{T}^\pi)^{-1} \mathbf{R}^\pi. \end{aligned}$$

Melhoria de Política

Próxima Política π':

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

Resultado:

$$V^{\pi}(s) \leqslant V^{\pi'}(s).$$

Iteração de Política

1. escolha uma política arbitrária π_0
2. faça para toda iteração $k \geqslant 0$
   (a) avalie a política atual π_k
   (b) obtenha uma política melhorada π_{k+1}

   enquanto existir $s \in \mathcal{S}$ tal que:

$$V^{\pi_{k+1}}(s) < \max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V^{\pi_{k+1}}(s') \right\}$$

3. retorne π_{k+1}

Exemplo

Considere um MDP com estados $\{s_0, s_1, s_2, \ldots, s_{N-1}, s_N = g\}$ tal que g seja um estado absorvedor. Considere um fator de desconto γ e a política que executa uma ação (custo c) que vai para o próximo estado em cada um dos seguintes casos abaixo:

1. Fica no mesmo estado com probabilidade p no estado s_{N-1}
2. Retorne para o começo com probabilidade p no estado s_{N-1}
3. Fica no mesmo estado com probabilidade p em todos os estados
4. Retorne para o começo com probabilidade p em todos os estados

Calcule o valor de s_0 em cada um dos casos.

Iteração de Política Modificado

Considere o seguinte operador:

$$(\mathcal{T}^\pi V)(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V(s').$$

Avaliação de Política (iteração de valor):

$$V^\pi = \lim_{n \to \infty} (\mathcal{T}^\pi)^n V,$$

para qualquer V arbitrário.

Iteração de Política Modificado

1. inicializa $V_0(s)$ arbitrariamente
2. faça para toda iteração $k \geqslant 0$
   (a) obtenha política π para todo $s \in \mathcal{S}$

$$\pi(s) = \arg\max_{a \in \mathcal{A}} \left\{ [R(s,a) + \sum_{s' \in \mathcal{S}} T(s,a,s') V_k(s') \right\}$$

   (b) repita m vezes para todo $s \in \mathcal{S}$

$$V_{k+1}(s) = R(s, \pi(s)) + \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V_k(s')$$

   (c) para todo $s \in \mathcal{S}$

$$V_{k+1}(s) = \max_{a \in \mathcal{A}} \left\{ R(s,a) + \sum_{s' \in \mathcal{S}} T(s,a,s') V_k(s') \right\}$$

   enquanto $\|V_k - V_{k+1}\|_\infty < \epsilon$
3. retorne a política

$$\pi(s) = \arg\max_{a \in \mathcal{A}} \left\{ R(s,a) + \sum_{s' \in \mathcal{S}} T(s,a,s') V_k(s') \right\}$$

Iteração de Política Modificado

- operador T^{π} exibe uma complexidade de $O(|\mathcal{S}|^2)$
- o operador T exibe uma complexidade de $O(|\mathcal{A}| \cdot |\mathcal{S}|^2)$
- a convergência do algoritmo Iteração de Valor ocorre primeiramente na política e depois na função valor

Shortest Sthocastic Path (sem desconto)

Definition 2. Uma política π é própria se $\lim_{t \to \infty} \Pr(s_t \in \mathcal{G} | \pi) = 1$.

Condição para existência de Política Ótima:

1. Existe pelo menos uma política própria.

2. Para toda política imprópria, existe um estado inicial que acumula custo infinito sob esta política.

O algoritmo de Iteração de valor converge.

O algoritmo de Iteração de Política converge se a política inicial π_0 for própria.

Programação Linear: Equação de Bellman

Considere que $b(s)$ são valores não-negativos e $b(s_0) > 0$

$$
\begin{aligned}
\underset{V(s)}{\text{minimize}} \quad & \sum_{s \in \mathcal{S} \setminus G} b(s) V(s) \\
\text{s.t.} \quad & V(s) \geqslant R(s,a) + \gamma \sum_{s' \in \mathcal{S}} T(s,a,s') V(s'), \quad \forall s \in \mathcal{S} \setminus \mathcal{G}, a \in A \ ; \\
& V(s) = 0, \qquad \forall s \in \mathcal{G}
\end{aligned}
$$

$$
\pi(s) = \arg\max_{a \in \mathcal{A}} \left\{ R(s,a) + \gamma \sum_{s \in \mathcal{S}} T(s,a,s') V_k(s') \right\}
$$

Programação Linear: Fluxo Constante

$$\begin{aligned} \underset{x(s,a)}{\text{maximize}} \quad & \sum_{s \in \mathcal{S} \setminus \mathcal{G}} \sum_{a \in \mathcal{A}} x(s,a) R(s,a) \\ \text{s.t.} \quad & \sum_{a \in \mathcal{A}} x(s,a) - \gamma \sum_{s' \in \mathcal{S}} T(s',a,s) x(s',a) = b(s), \quad \forall s \in \mathcal{S} \setminus \mathcal{G} \; ; \\ & x(s,a) \geqslant 0, \qquad \forall s \in \mathcal{S} \setminus \mathcal{G} \end{aligned}$$

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

Relatório 1

Estudo empírico na complexidade de diversos algoritmos para resolver MDP:

- Iteração de Valor (síncrono, assíncrono, ϵ)
- Iteração de Política (iteração de valor)
- Iteração de Política Modificado (parâmetro m)
- Pelo menos um algoritmo das próximas aulas

Será disponibilizado um problema enumerado de navegação robótica.

Será disponibilizado um problema em linguagem RDDL.

Mas, pode-se considerar outros ambientes para demonstrar alguma propriedade específica.

Problema, formato e limite de páginas especificado no TIDIA.

Resumos

Conteúdo (2 páginas, formato no TIDIA):

- *Abstract* original
- Problema específico que se deseja resolver
- Contribuição do Artigo
- Experimentos (aplicação, problema, comparação)