

# Planejamento Probabilístico

Helton Massato Kishi

14 de novembro de 2007

# Conteúdo

- O que é Planejamento Probabilístico?
- Algoritmo de iteração por valor
- Algoritmo RTDP
- Exemplo

# O que é planejamento probabilístico?

- Planejamento é o processo de escolher e organizar ações através da antecipação de seus efeitos, tendo como objetivo satisfazer uma meta pré-estabelecida.
- Planejamento clássico:
  - Conjunto de estados, que são descrições do sistema.
  - Conjunto de ações, que são as ações que o agente pode executar em um dado momento.
  - Quando executamos uma ação em um estado, sabemos com total certeza qual será o próximo estado.
  - O problema de planejamento clássico é achar um conjunto de ações que levam o sistema de um dado estado inicial para um estado final.
- Planejamento Probabilístico:
  - Os efeitos das ações são incertos.
  - Quando executamos uma ação em um estado, só sabemos a distribuição de probabilidade de qual será o próximo estado.

# Exemplo: Robô de entrega de cartas

- Robô que faz a entrega de cartas dentro de um andar de escritórios composto por três salas:
  - Sala do chefe
  - Sala do funcionário
  - Sala da secretária.
- O robô deve decidir, em um dado instante, para quem deve entregar a carta.
- Nem sempre há sucesso na entrega das cartas: existe a possibilidade de o destinatário não estar presente na sala quando o robô tenta fazer uma entrega.
- Prioridades diferentes. É muito mais importante entregar a carta ao chefe do que entregar a carta para a secretária.

# MDP - Markov Decision Process

- Modelo matemático usado para resolver o problema de planejamento probabilístico.
- Um MDP pode ser definido pela tupla  $\langle S, A, P, R, C \rangle$ , sendo:
  - $S$ , um conjunto de estados;
  - $A$ , um conjunto de ações;
  - $P : A \times S \times S \rightarrow [0, 1]$  uma função de transição de estados, onde  $P(a, s, s')$  é a probabilidade do agente ir do estado  $s$  para o estado  $s'$  após da execução da ação  $a$ ;
  - $R : S \rightarrow \mathbf{R}$  uma função de recompensa.  $R(s)$  é a recompensa se o sistema chegar no estado  $s$ .
  - $C : S \times A \rightarrow \mathbf{R}$ , uma função do custo.  $C(s, a)$  é o custo da ação  $a \in A$  no estado  $s \in S$ .
- A solução do MDP é uma *política*.
- Uma *política*  $\pi$  é uma função  $\pi : S \rightarrow A$ , em que mapeia para cada estado, qual é a melhor ação que deve ser tomada.

# Como escolher uma política?

- *História*: sequência de estados, ações e observações geradas desde o estado 0 até um certo instante de tempo de interesse.
- *Função de Valor  $V^*$* : Mapeia histórias em  $\mathbf{R}$ . Tem o objetivo calcular a qualidade das histórias. Uma história  $h$  é melhor do que  $h'$  se  $V^*(h) < V^*(h')$ .
- Função de valor para um horizonte  $T$ :

$$V^*(h) = \sum_{t=0}^{T-1} \{C(s_t, a_t) - R(s_t)\} - R(s_T).$$

- Função de valor para Horizonte infinito:

$$V^*(h) = \sum_{t=0}^{\infty} (\gamma^t (C(s_t, a_t)) - R(s_t))$$

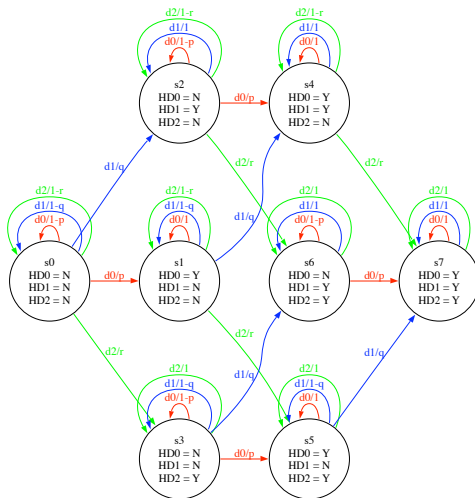
onde  $\gamma$  é a taxa de desconto ( $0 \leq \gamma < 1$ ). Quanto menor o  $\gamma$ , menor será a importância dos acontecimentos longe do início.

# Exemplo: Robô de entrega de cartas

- O estado pode ser definido usando três variáveis booleanas:
  - *HD0*: Verdadeiro se a carta para a secretária já foi entregue.
  - *HD1*: Verdadeiro se a carta para o funcionário já foi entregue.
  - *HD2*: Verdadeiro se a carta para o chefe já foi entregue.
- O conjunto de estados deste exemplo são todas as combinações possíveis das variáveis booleanas *HD0*, *HD1*, *HD2*, totalizando  $2^3 = 8$  estados.
- Conjunto de ações:
  - *d0*: entregar uma carta para a secretária
  - *d1*: entregar uma carta para o funcionário
  - *d2*: entregar uma carta para o chefe
- Probabilidades:
  - *p*: prob. de encontrarmos a secretária em sua sala = 0,9
  - *q*: prob. de encontrarmos o funcionário em sua sala = 0,7
  - *r*: prob. de encontrarmos o chefe em sua sala = 0,5

# Exemplo: Robô de entrega de cartas

Grafo da função de transição do problema:





# Exemplo: Robô de entrega de cartas

- Custo da ação: energia gasta pelo robô para entregar a carta, que é a mesma para todos os destinatários. Neste exemplo,  $C(s) = 10, \forall s \in S$ .
- A recompensa é definida da seguinte maneira:
- $R(s) = HD0 * 1 + HD1 * 2 + HD2 * 4$
- $R[] = [0, 1, 2, 4, 3, 5, 6, 7]$

# Algoritmo de Iteração por Valor (IV)

- Encontra uma política ótima  $\pi : S \rightarrow A$ .
- Para encontrarmos tal política é preciso escolher ações que minimizem a função de valor para horizonte infinito. Isto é calculado usando a equação de **Bellman**, definida abaixo:

$$V(s) = \min_{a \in A} \{ C(a, s) - R(s) + \sum_{s' \in S} \gamma^n P(a, s, s') V(s') \}$$

A política pode ser extraída pela equação:

$$\pi(s) = \arg \min_{a \in A} \{ C(a, s) - R(s) + \sum_{s' \in S} \gamma^n P(a, s, s') V(s') \}$$

# Algoritmo de Iteração por Valor (IV)

Iteração por valor( $S, A, P, R, C, \epsilon$ )

- Inicializa um vetor  $V(s)$  com valores arbitrários.
- Enquanto  $\exists s \in S$  tal que  $Residual(s) > \epsilon$  faça:
  - Atualiza valores de  $V(s)$  **para todo**  $s \in S$ , usando a equação de Bellman.
  - Cálculo de  $Residual(s)$  para todo  $s \in S$ , que é a diferença entre os valores de  $V(s)$  antes e depois da iteração.

# Política encontrada por IV para o problema do robô

Política IV	
Estado	Ação
s0	d2
s1	d2
s2	d2
s3	d1
s4	d2
s5	d1
s6	d0
s7	d0

# SSP: Caminho Estocástico Mínimo

- Sub-problema do MDP.
- É necessário definir o estado inicial  $s_0$  e o conjunto dos estados meta  $S_m$ .
- Assim, um problema de planejamento probabilístico pode ser naturalmente descrito como um problema de SSP.

# Algoritmo RTDP

Resolve o problema de SSP, devolvendo uma política parcial.

Algoritmo RTDP( $S, A, P, R, C, s_0, S_m$ )

- Inicializa os valores de  $V(s)$ .
- $s = s_0$ ; //  $s$  é o estado atual
- Enquanto  $s \notin S_m$  faça:
  - Recalcule  $V(s)$  usando a equação de Bellman.
  - Simule a execução da melhor ação dada por  $V(s)$ .
  - Atualiza o valor de  $s$ .

# Política encontrada por RTDP para o problema do robô

Política RTDP	
Estado	Ação
s0	d0
s1	d1
s2	null
s3	null
s4	d2
s5	null
s6	null
s7	null

# Comparação entre IV e RTDP

- IV encontra política *ótima*, enquanto RTDP encontra política *parcial*.
- O número de atualizações de  $V(s)$  é muito menor no RTDP. As atualizações são feitas apenas nos estados que são percorridos pela simulação.