Proposta de Trabalho:
MDP e Equação de Bellman

Rodrigo Luna

September 19, 2025

Objetivos do Trabalho

- Definir claramente um **Processo de Decisão de Markov (MDP)**.
- Escolher um dos exemplos de problemas Toys descrito no gymnasium: gymnasium
- Aplicar a **Equação de Bellman** para calcular:
  - Função de valor dos estados, $V(s)$.
  - Função de valor estado-ação, $Q(s, a)$.
- Interpretar os resultados obtidos.

Estrutura do Trabalho

1. Definição do MDP:
   - Estados ($\mathcal{S}$)
   - Ações ($\mathcal{A}$)
   - Probabilidades de Transição ($P(s' \mid s, a)$)
   - Função de Recompensa ($R(s, a)$)
   - Fator de Desconto (γ)
2. Cálculo iterativo da Equação de Bellman para encontrar $V(s)$ e $Q(s, a)$.
3. Aplicação do Value Iteration
4. Aplicação do Policy Iteration
5. Discussão e interpretação dos resultados.

Exemplo Prático: MDP mais complexo

Considere um Grid 3×3:

- Estados: $\mathcal{S} = \{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$
- Ações: $\mathcal{A} = \{Cima, Baixo, Esquerda, Direita\}$
- Recompensas:
  - $R(s,a) = -1$ por passo.
  - Estado $(3,3)$ é terminal com recompensa $+20$.
  - Estado $(2,2)$ tem recompensa negativa -10.
- Transição: 80% para o estado indicado pela ação, 10% para estados laterais (ortogonais), senão permanece no mesmo estado.
- Fator de desconto: $\gamma = 0.9$

Cálculo de $V(s)$

Use a Equação de Bellman iterativamente:

$$V(s) = \max_a \left\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V(s') \right\}$$

Exemplo (para o estado $(2,1)$):

$$\begin{aligned} V((2,1)) = \max\big\{ & -1 + 0.9[0.8V((1,1)) + 0.1V((2,2)) + 0.1V((3,1))], \\ & -1 + 0.9[0.8V((3,1)) + 0.1V((2,2)) + 0.1V((1,1))], \\ & -1 + 0.9[0.8V((2,1)) + 0.1V((1,1)) + 0.1V((3,1))], \\ & -1 + 0.9[0.8V((2,2)) + 0.1V((3,1)) + 0.1V((1,1))]\big\} \end{aligned}$$

Calcule iterativamente até convergência.

Cálculo de $Q(s,a)$

A função valor-ação é dada por:

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V(s')$$

Exemplo (estado $(2,1)$, ação $Cima$):

$$Q((2,1), Cima) = -1 + 0.9[0.8V((1,1)) + 0.1V((2,2)) + 0.1V((3,1))]$$

Repita para todas as ações em todos os estados.

Orientações Adicionais

- Grupos com no máximo **3 alunos**.
- Relatório claro, contendo definições, cálculos detalhados e interpretações.
- Sugere-se utilização de Python, código explicado em sala de aula.
- Apresentação breve dos resultados principais em aula.

Avaliação do Trabalho

- Clareza e precisão na definição do MDP.
- Correção e detalhamento dos cálculos.
- Compreensão e interpretação correta dos resultados.
- Organização e apresentação do relatório.