MAEXXXX - Processos de Decisão de Markov e Aprendizado por Reforço

Prof. L. Felipe Martins
Department of Mathematics and Statistics
Cleveland State University
l.martins@csuohio.edu

1 Objetivo

Proporcionar ao estudante uma formação sólida em processos de decisão sequenciais, com ênfase em Processos de Decisão de Markov e Aprendizado por Reforço. Ao final da disciplina, espera-se que o estudante seja capaz de modelar problemas de decisão sob incerteza, compreender os fundamentos teóricos das funções de valor e das equações de Bellman, implementar algoritmos clássicos de programação dinâmica e métodos modernos de aprendizado por reforço, bem como analisar criticamente seu desempenho em problemas de interesse prático.

2 Justificativa

Processos de decisão sequenciais constituem um dos principais paradigmas da Inteligência Artificial moderna, com aplicações em robótica, sistemas autônomos, finanças, saúde, controle, logística e sistemas de recomendação. O Aprendizado por Reforço, fundamentado na teoria dos Processos de Decisão de Markov, tornou-se uma das áreas mais importantes da ciência de dados e da inteligência artificial, impulsionado por avanços recentes em aprendizado profundo e poder computacional. A disciplina fornece ao estudante os fundamentos matemáticos, computacionais e algorítmicos necessários para compreender, desenvolver e aplicar métodos de otimização e aprendizado em problemas de decisão sob incerteza, contribuindo para sua formação em pesquisa e para o desenvolvimento de aplicações em diferentes áreas do conhecimento.

3 Ementa

Neste curso, são estudados *processos de decisões sequenciais*, em que um *agente* atua num *ambiente*, e recebe *recompensas*, que são uma função tanto das ações selecionadas como do estado em que o ambiente se encontra. Exemplos de áreas de aplicação incluem robótica, jogos, sistemas de recomendação, saúde e controle.

O modelo matemático usado para este tipo de problema é chamado *Processo de Decisão de Markov*. Neste modelo, uma vez que o agente seleciona uma política, que determina (possivelmente de forma probabilística) as ações a serem tomadas em cada estado do ambiente, o processo segue

a dinâmica de um *processo de Markov com recompensas*, que possui uma estrutura amplamente estudada. O objetivo é maximizar o total das recompensas a longo prazo.

No curso, são estudados métodos tabulares clássicos (programação dinâmica) e técnicas modernas (aprendizado por reforço). O curso abrange tanto aspectos teóricos como práticos para a criação de algoritmos de optimização. Para experimentação, são usadas ferramentas "open-source", incluindo Python, NumPy, Jupyter, Gymnasium and PyTorch.

4 Pré-requisitos

Probabilidade, álgebra linear.

O curso não assume conhecimento prévio de programação, mas familiaridade com alguma linguagem de programação é recomendada.

5 Avaliação

Instrumentos de avaliação no curso incluem listas de exercícios, laboratórios computacionais e um projeto final apresentado em classe.

6 Referências Bibliográficas

1. Sutton and Barto, *Reinforcement Learning: An Introduction*
   `http://incompleteideas.net/book/RLbook2020.pdf`
2. Powell, *Reinforcement Learning and Stochastic Optimization*
   `https://castle.princeton.edu/wp-content/uploads/2019/10/Powell-Reinforcement-Learning-a pdf`
3. Bertsekas, *A Course in Reinforcement Learning*
   `https://www.mit.edu/~dimitrib/RLCOURSECOMPLETE%202ndEDITION.pdf`
4. Szepesvári, *Algorithms for Reinforcement Learning*
   `https://sites.ualberta.ca/~szepesva/papers/RLAlgsInMDPs.pdf`
5. Plaat, *Deep Reinforcement Learning*
   `https://deep-reinforcement-learning.net`
6. Cai, *An Introduction to Reinforcement Learning*
   `https://rlbook.adzc.ai`
7. Xiao, *Reinforcement Learning: Theory and Python Implementation*
   `https://zhiqingxiao.github.io/rl-book`

7 Tópicos

1. Processos de Decisão de Markov (PDM/MDP): definição de ambiente, estado, agente, ações e recompensa. O ciclo evolutivo de um PDM. Exemplos de PDMs. Exemplos de modelagem. Introdução às ferramentas computacionais.

2. Função de valor e equação de Bellman: função valor de estado e função valor de ação. Equações de Bellman para uma política específica. Métodos numéricos exatos e iterativos. Propriedade de contração, existência de soluções e convergência.

3. Função de valor optimal e equação de Bellman optimal e o Teorema de Verificação. Métodos de solução: iteração de valor e iteração de política. Implementação prática em modelos simples.

4. Introdução a aprendizado sem modelo. Problemas de "bandits" e o conflito exploração–aproveitamento. Estratégias ε-greedy. Limites superiores de confiança e amostragem de Thompson. Experimentação com estratégias ε-greedy.

5. Métodos de Monte-Carlo: Predição e controle. Métodos "on-policy" e "off-policy". Aplicação a problemas de complexidade moderada, como Blackjack.

6. Diferenças temporais (TD). TD(0), SARSA e "Q-Learning". Métodos de múltiplos passos, traços de eligibilidade e TD(λ).

7. Aproximação de funções. Limitações de métodos tabulares: problemas com modelos definidos e problemas livres de modelo. Representações por atributos, aproximação linear, métodos de semigradiente. Métodos de gradiente de política: REINFORCE e estimadores baseados em gradiente.

8. Métodos ator-crítico. Funções de vantagem, redução de variância, arquiteturas ator-crítico.

9. Aprendizado profundo por reforço. Introdução a redes neurais, "Q-learning" profundo, repetição de experiências, redes-alvo, extensões modernas.