HW 5_6000

Task 1:

Markov Decision Process (MDP) is a foundational element of reinforcement learning (RL). MDP allows formalization of sequential decision making where actions from a state not just influences the immediate reward but also the subsequent state.

In an MDP, an agent interacts with an environment by taking actions and seek to maximize the rewards the agent gets from the environment. At any given timestamp t, the process is as follows

- The environment is in state S_t
- The agent takes an action A_t
- The environment generates a reward R_t based on S_t and A_t
- The environment moves to the next state S_{t+1}

Goal is to maximize total rewards (ΣR_t) collected over a period of time.

A hypothetical construct of a real-world application that can be formulated as MDP-

- Title: Quiz Game Show
- Application/Goal: In a quiz game show there are 10 levels, at each level one question is asked and if answered correctly a certain monetary reward based on the current level is given. Higher the level, tougher the question but higher the reward. At each round of play, if the participant answers the quiz correctly then s/he wins the reward and gets to decide whether to play at the next level or quit. If quit, then the participant gets to keep all the rewards earned so far. At any round if participants failed to answer correctly then s/he loses "all" the rewards earned so far. The game stops at level 10. The goal is to decide on the actions to play or quit maximizing total rewards
- States: {level1, level2, …, level10}
- Actions: Play at next level or quit
- State Transitions: State transitions are uni-directional and probabilistic in nature. A player cannot go back to a previous state; s/he whether proceeds to the next state or quits from current state. The higher the state, the higher the probability that player will quit based on the toughness of the question asked.
- Rewards: Play at level1, level2, …, level10 generates rewards \$10, \$50, \$100, \$500, \$1000, \$5000, \$10000, \$50000, \$100000, \$500000 with probability p = 0.99, 0.9, 0.8, …, 0.2, 0.1 respectively. The probability here is the probability of giving correct answer in that level. At any level, the participant losses with probability (1- p) and losses all the rewards earned so far.

Task 2:

Sector: Healthcare

Project Title: Deep Reinforcement Learning for Sepsis Treatment

Open-Source Project Link: https://github.com/aniruddhraghu/sepsisrl

Problem: Sepsis is a leading cause of mortality in intensive care units and costs hospitals billions annually. Treating a septic patient is highly challenging because individual patients respond very differently to medical interventions and there is no universally agreed-upon treatment for sepsis.

Solution Using RL: Authors propose an approach to deduce treatment policies for septic patients by using continuous state-space models and deep reinforcement learning. Their model learns clinically interpretable treatment policies, similar in important aspects to the treatment policies of physicians. The learned policies could be used to aid intensive care clinicians in medical decision making and improve the likelihood of patient survival.

Details:

This article was one of the first ones to directly discuss the application of deep reinforcement learning to healthcare problems.

Data: In the article the authors use the Sepsis subset of the MIMIC-III (Multiparameter Intelligent Monitoring in Intensive Care) dataset. Data are aggregated into windows of 4 hours This yielded a 48 × 1 feature vector for each patient at each timestep.

Actions and Rewards: They choose to define the action space as consisting of Vasopressors and IV fluid. They group drugs doses into four bins consisting of varying amounts of each drug. They discretized the action space into per-drug quartiles based on all non-zero dosages of the two drugs and converted each drug at every timestep into an integer representing its quartile bin. The reward function is clinically motivated based on the SOFA score which measures organ failure. Our reward function penalizes high SOFA scores and increases in SOFA score and lactate levels. Positive rewards are issued for decreases in these metrics.

Network Architecture: The core network is a Double-Deep Q Network which has separate value and advantage streams.

Results: For evaluation they use what Gottesman termed as the U-Curve. Specifically, they look at the mortality rate as a function of the difference in dosage of the prescribed policy versus the actual policy.

References:

- https://www.mygreatlearning.com/blog/reinforcement-learning-in-healthcare/
- https://towardsdatascience.com/real-world-applications-of-markov-decision-process-mdp-a39685546026
- https://towardsdatascience.com/a-review-of-recent-reinforcment-learning-applications-to-healthcare-1f8357600407
- https://arxiv.org/abs/1711.09602
- https://www.capestart.com/resources/blog/reinforcement-learning-in-health-care-why-its-important-and-how-it-can-help/
- https://www.jmir.org/2020/7/e18477/