Breunna Bingham

Southern New Hampshire University

June 5, 2023

CS-370 Current/Emerging Trends in CS

Module Six Assignment: Cartpole Revisited

The Cartpole problem can be solved using the reinforce algorithm. The goal is to discover a policy that maximizes the total reward that will be received from any time t until the terminal time T (Yoon, 2021a). This allows for the algorithm to choose the best possible choice when making decisions as it learns from trial and error over time and experience. The more experience the AI has, the more successful the algorithm becomes at making correct decisions. Input is taken in and then an action is taken based on the current state of the agent. Feedback can be positive or negative in this case and that is where the algorithm learns the policy by mapping the reward to the state. It looks a little something like this diagram:

A picture containing text, diagram, line, screenshot

Description automatically generated(Bhatt, 2019)

The cartpole problem can also be solved using the Advantage Actor-Critic or A2C algorithm. The Advantage Actor Critic (A2C) algorithm in the field of reinforcement learning combines two types of reinforcement learning algorithms (Policy Based and Value Based) together. Policy Based agents learn a policy (a probability distribution of actions) directly mapping input states to output actions. Value Based algorithms learn to select actions based on the predicted value of the input state or action (Wang, 2023). The main advantage of using A2C is that it can solve a wider range of problems with a lower variance in performance when compared to the reinforce algorithm (*PyLessons*, 2020).

A picture containing diagram, text, sketch, line

Description automatically generated

(Yoon, 2021b)

A policy gradient approach differs from a value-based approach. A policy gradient uses the probability of the distribution of actions so that the actions with a higher reward also have a higher probability for a state (Yoon, 2021a). The goal for the policy gradient is to find the actions that give out the highest rewards. Value-based algorithms uses the input state to predict a value and select actions based upon that predicted value. The goal is to find the maximum value.

Actor-Critic uses both value and policy algorithms in its algorithm and combines them into a single step. Actor-Critic algorithm differs from both the policy and value approaches by making a decision based upon a single step. For example, you have an actor, and a critic. The actor is playing the game but the critic is advising on if the action the actor takes is a good or bad action etc. So, each action that is taken, it is giving feedback and that feedback is reliant on a single action that the algorithm learns from. With policy and value-based algorithms, both use previous multiple actions to formulate their probability of making the next action a ‘correct’ action.

Resources

Bhatt, S. (2019, April 19). Reinforcement Learning 101 - Towards Data Science. *Medium*. <https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292>

*PyLessons*. (2020, March 20). <https://pylessons.com/A2C-reinforcement-learning>

Wang, M. (2023, June 3). Advantage Actor Critic Tutorial: minA2C - Towards Data Science. *Medium*. https://towardsdatascience.com/advantage-actor-critic-tutorial-mina2c-7a3249962fc8#:~:text=In%20the%20field%20of%20Reinforcement,input%20states%20to%20output%20actions.

Yoon, C. (2021, December 7). Deriving Policy Gradients and Implementing REINFORCE. *Medium*. https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63

Yoon, C. (2021b, December 7). Understanding Actor Critic Methods and A2C - Towards Data Science. *Medium*. https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f