Assignment
Deep Reinforcement Learning

Akash Sharma
UBIT-as475
as475@buffalo.edu
Department of Computer Science and Engineering
University at Buffalo
Buffalo, NY 14214

10 December 2021

1 Introduction

Goal in this Assignment to learn the trends in stock price and perform a series of trades over a period of time and end with a profit with the help of Reinforcement Learning of the model.

1.1 What is Reinforcement Learning?

Reinforcement Learning is training the model based on rewarding desired behaviours and punishing the undesired behaviours. In simple words, agent in model is able to perceive and interpret the environment and take action according to the learning through trial and error.

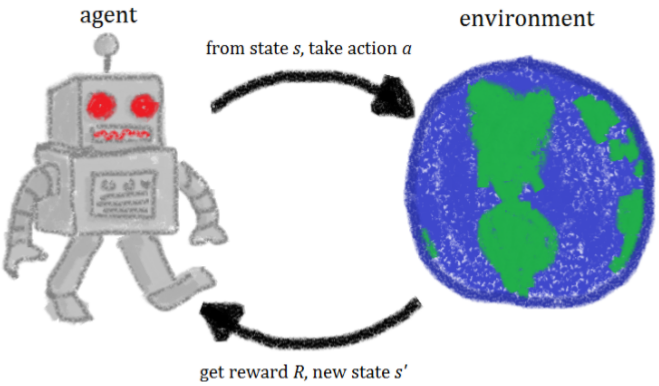
1.2 Data set and its features

Dataset consist of historical stock price for Nvidia for the last 5 years. The dataset has 1258 entries starting 10/27/2016 to 10/26/2021. The features include information such as the price at which the stock opened, the intraday high and low, the price at which the stock closed, the adjusted closing price and the volume of shares traded for the day.

2 Explanation of Stock Trading Environment

2.1 What is Environment?

Environment is very fundamental element of Reinforcement Learning. It help us to take right decision for the agent. When agent interact with environment, it returns a new state and rewards for the action and store the value in Qtable for the future use.


agent
environment
from state s, take action a
get reward R, new state s'

Above picture is fictional representation of the environment.

2.2 Stock Trading Environment

Stock Trading environment reading the data from the NVDA.csv file which consist of stock price of Nvidia of last 5 years and features includes the price at which the stock opened, the intraday high and low, the price at which the stock closed, the adjusted closing price and the volume of shares traded for the day.

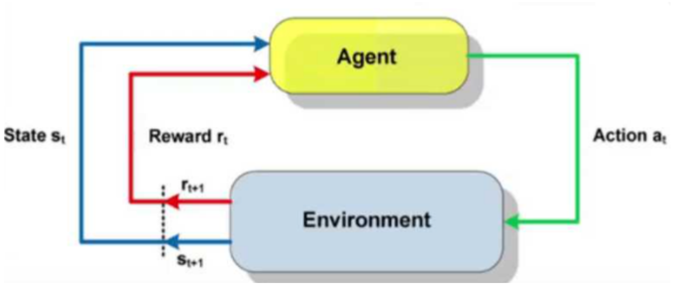
The environment calculates the trend of stock market of Nvidia. It has 4 states and 3 actions. Actions consist of Buy, Sell and Hold and having a integer range 0 to 2 respectively. In the program step method called to agent to take action. Step method returns 4 values which are observation, rewards, done and info.
1. Observation - There are 4 possible observation which agent receive in stock Trading environment, it depends on price increased on average no of days and also agent consider whether agent already stock or not.
2.Rewards - These are values which is use to measure performance of agent.
3.Done - Theses are boolean which is use to describe whether episode is complete or not. 4.Info - It is information if the implementation.

3 Implementation of Q-Learning

Q learning is off policy reinforcement learning algorithm that find the best action for the current state.It is called off policy because it take random action and any predefined policy is not needed. Q in Q learning is stand for the quality which means how useful is current data for the future rewards.

3.1 Steps to Implement Q Learning

3.1.1 Architecture of Qlearning



3.1.2 Create and Making a updates in Qtable

Qtable is a matrix which consist of state and action. Qtable is firstly initialised with the zero. We update the qtable with every episode which is nothing but a iteration. This qtable help agent to select the best option .
Agent simply interact with environment and make update in the qtable. There are two ways to interact with the environment.
1. Exploration- When the agent select the action randomly is called exploration.

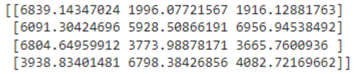
2. Exploitation- When the agent select the action based on maximum value that selection is called exploitation.

Both exploration and exploitation has its own significance. Random selection is important because it help the agent to discover new states otherwise all states may not be selected. Exploration and exploitation can be balance by epsilon, basically epsilon decide the how often to explore or exploit.

3.1.3 Update Qtable

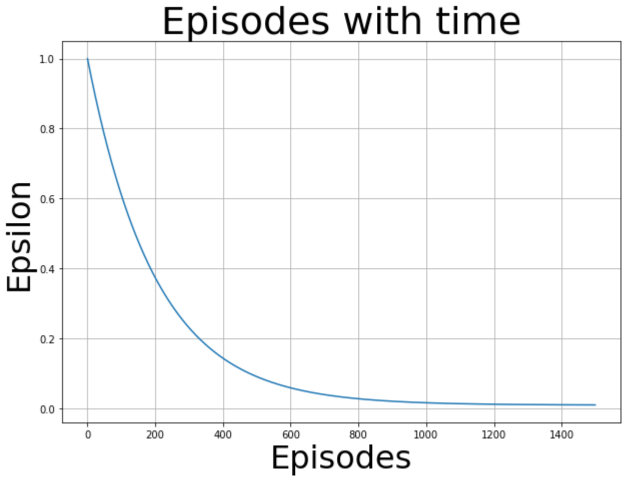
Updating of qtable occurs when episode is done. In every episode state and action is updated.Done means when agent meets the destination. Agent learns with every episode. There are below basics steps:
1.Agent starts with some state, take action and receive award.
2.Agent choose the action by refering Qtable with having highest value or random value and the update the qtable value.

3.1.4 Updated Qtable

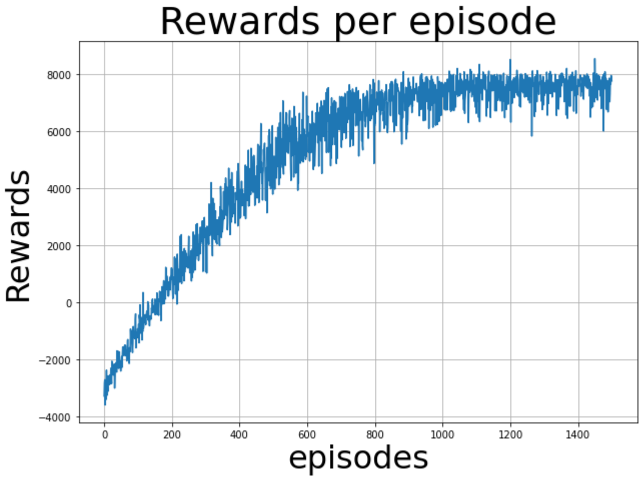


```
[[6839.14347024 1996.07721567 1916.12881763]
 [6091.30424696 5928.50866191 6956.94538492]
 [6804.64959912 3773.98878171 3665.7600936 ]
 [3938.83401481 6798.38426856 4082.72169662]]
```

4 Graphs for epsilon decay and total reward per episode



Above graph reprensents episodes with increasing time with respect to decreading value of epsilon which is nothing but epsilon decay.



Above graph represents rewards per episode.It is increasing with time since model is learning with time.

5 References

1. https://towardsdatascience.com/simple-reinforcement-learning-q-learning-fcddc4b6fe56

2. https://www.youtube.com/watch?v=WLOUElrYvyo

3. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.close.html