Assignment 3

Sanyam Gupta (50359154)

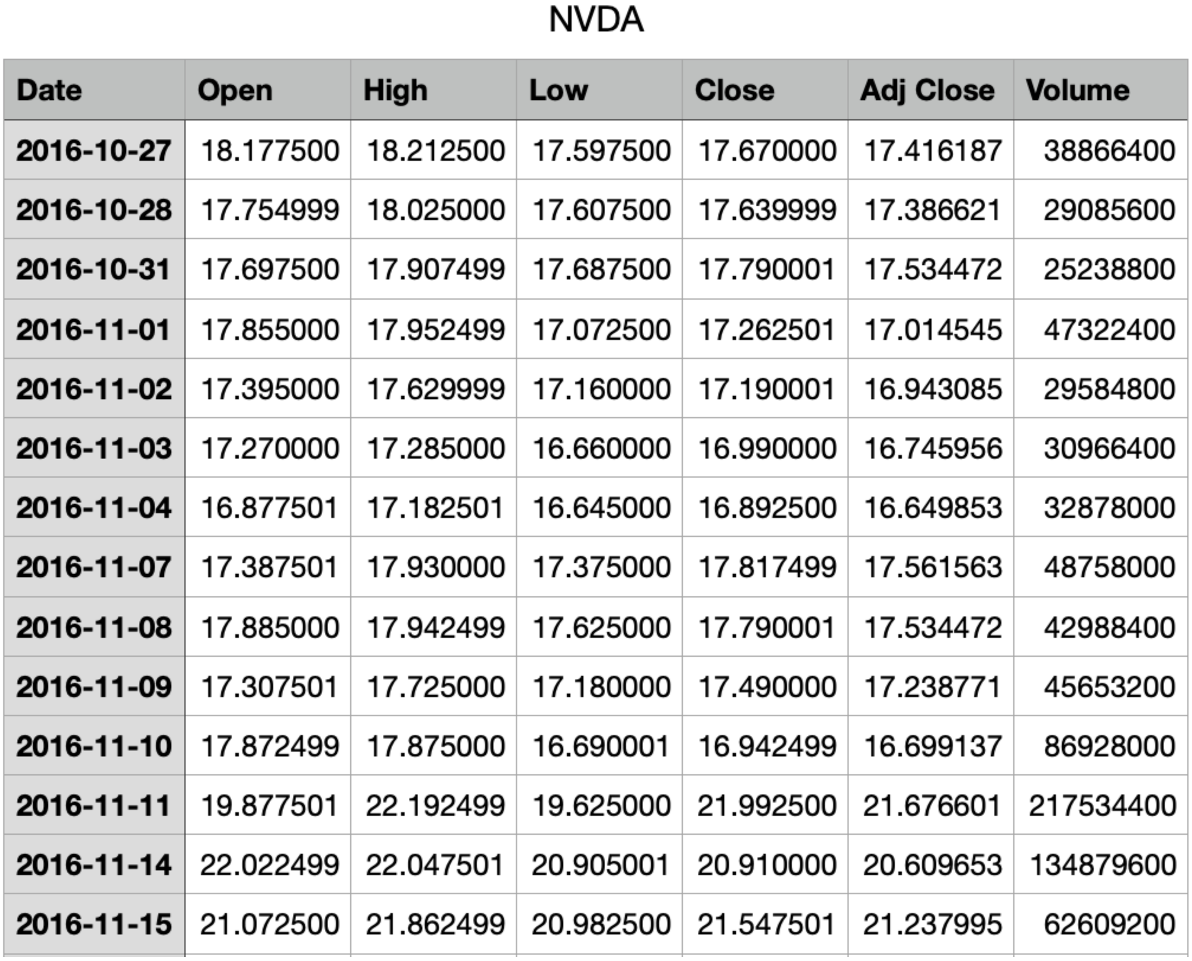
Dec 10 2021

1. Overview

The goal here is to use Reinforcement learning to learn trends in the stock market. After the agent has learnt, use it to perform a series of trades. Q learning is used here for Reinforcement learning.

2. Dataset

The dataset has stock information for NVDA stock price from 10/27/2016 to 10/26/2021.



NVDA

Date	Open	High	Low	Close	Adj Close	Volume
2016-10-27	18.177500	18.212500	17.597500	17.670000	17.416187	38866400
2016-10-28	17.754999	18.025000	17.607500	17.639999	17.386621	29085600
2016-10-31	17.697500	17.907499	17.687500	17.790001	17.534472	25238800
2016-11-01	17.855000	17.952499	17.072500	17.262501	17.014545	47322400
2016-11-02	17.395000	17.629999	17.160000	17.190001	16.943085	29584800
2016-11-03	17.270000	17.285000	16.660000	16.990000	16.745956	30966400
2016-11-04	16.877501	17.182501	16.645000	16.892500	16.649853	32878000
2016-11-07	17.387501	17.930000	17.375000	17.817499	17.561563	48758000
2016-11-08	17.885000	17.942499	17.625000	17.790001	17.534472	42988400
2016-11-09	17.307501	17.725000	17.180000	17.490000	17.238771	45653200
2016-11-10	17.872499	17.875000	16.690001	16.942499	16.699137	86928000
2016-11-11	19.877501	22.192499	19.625000	21.992500	21.676601	217534400
2016-11-14	22.022499	22.047501	20.905001	20.910000	20.609653	134879600
2016-11-15	21.072500	21.862499	20.982500	21.547501	21.237995	62609200

Fig 1 - Sample images in dataset

The information in each column is :

- Open - the price at which the stock opened.
- High - the intraday high.
- Low - the intraday low.

- Close - the price at which the stock closed.
- Adj Close - the adjusted closing price.
- Volume - the volume of shares traded for the day.

3. Environment Details

Actions, states, rewards and Goal are a function of the environment.

- Actions - Can be 0, 1, 2. Representing Buy, Sell, Hold.
- States - Can be 0, 1, 2, 3. Which are encoding of the following arrays. Each entry at index is a boolean value i.e 0 - Price Increased , 1 - Price Decreased, 2 - Stock Held, 3 - Stock Not held
  - 0 - [1, 0, 0, 1]
  - 1 - [1, 0, 1, 0]
  - 2 - [0, 1, 0, 1]
  - 3 - [0, 1, 1, 0]
- Goal is a flag Done
- Reward is a return of an action taken corresponding to an observation.

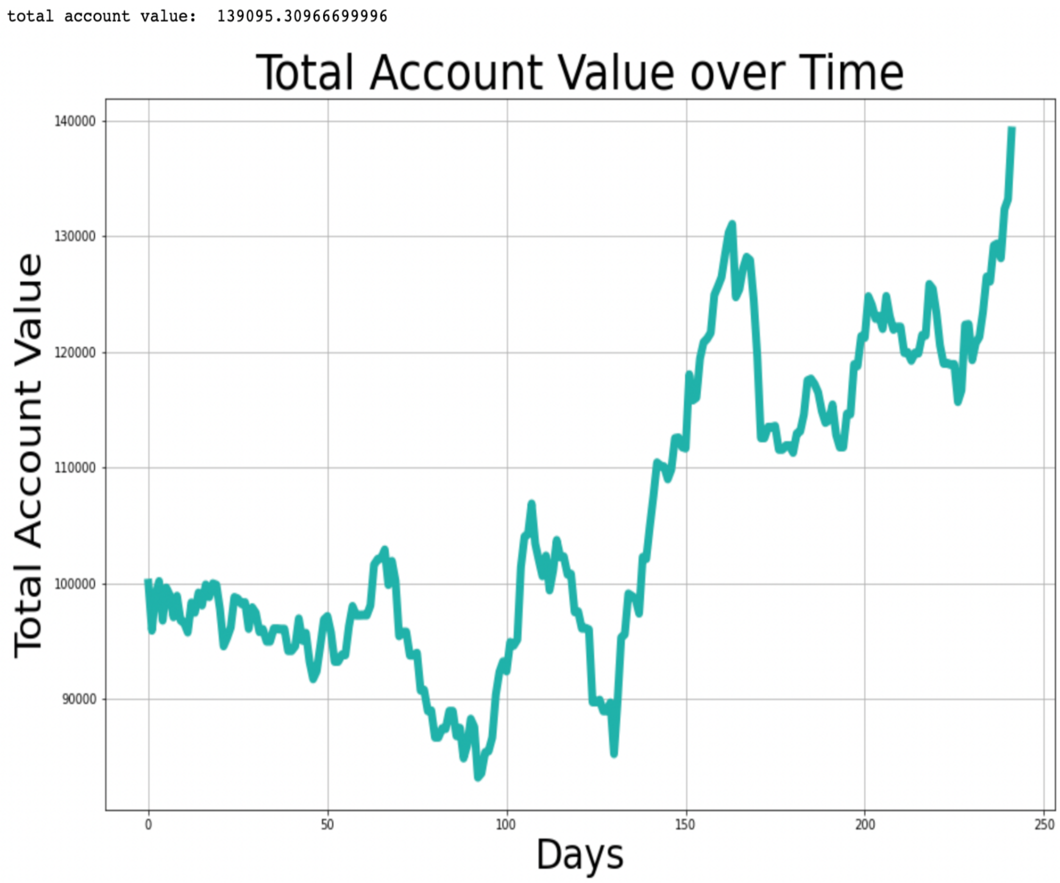
4. Q learning

4.1 Training and Testing the model

Deterministic Q learning algorithm is implemented from scratch and used as an RL algorithm. The Qtable is a 3 x 4 table. Where 3 is number of actions and 4 is number of observations. ε greedy selection method is used to decide which action to choose. During the training phase this value favours exploration during initial episodes and exploitation during later episodes. However, during testing a greedy selection method is used which always exploits.

4.2. Outcome

Starting from an account value of $100000 the trained agent increases the total value to **$139095.30**.



```
total account value:  139095.30966699996
```

Fig 2 - Model Performance

Explanation -

The graph value depicts account values over time. While selecting action we choose the most optimal action i.e the most greedy action at a state.

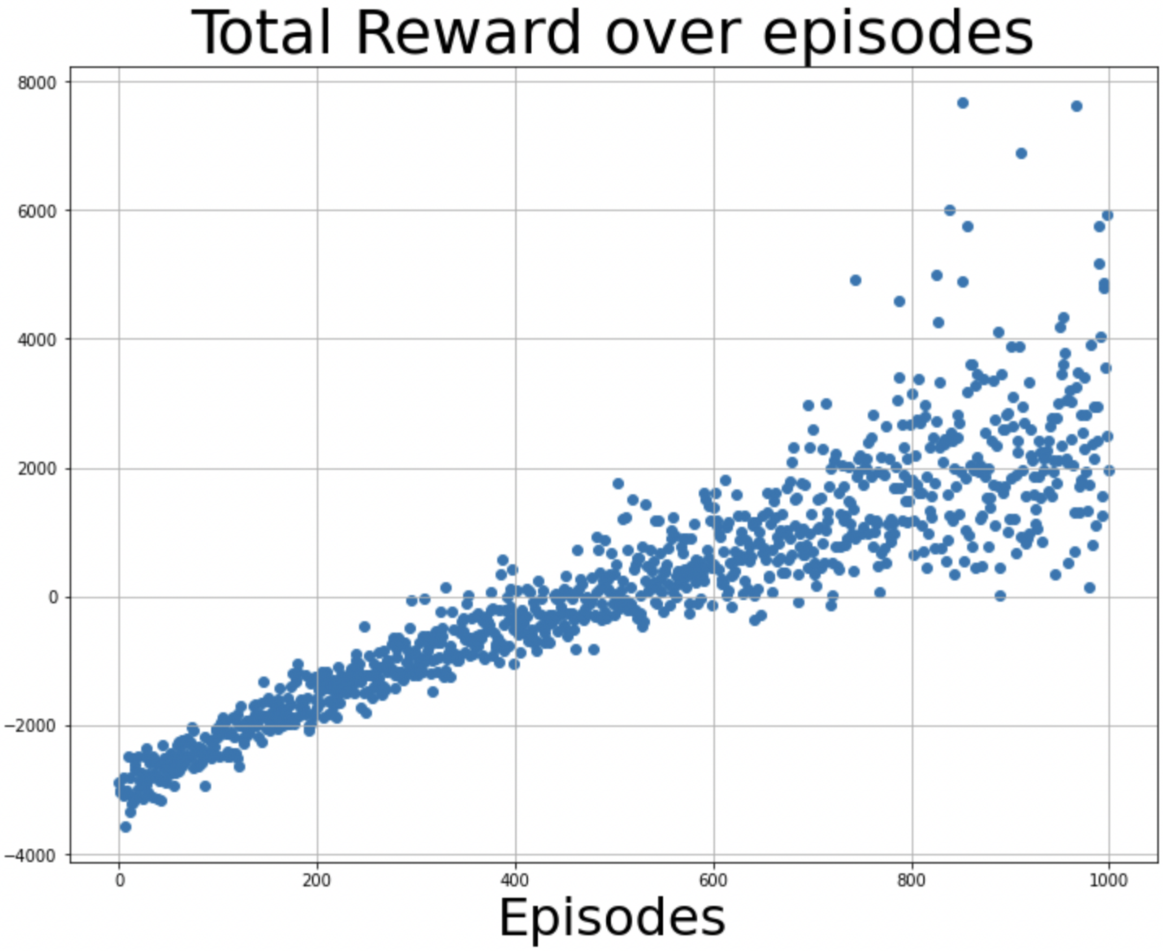


Fig 3 - Total rewards over Episodes

Explanation -
Since we start from all 0 values in Q tables. Rewards have large negative values in the initial phases. As the algorithm progresses the reward value becomes positive. Another factor contributing to large positive values is decrease in ε ε i.e we select more and more greedily as iterations increase.

-

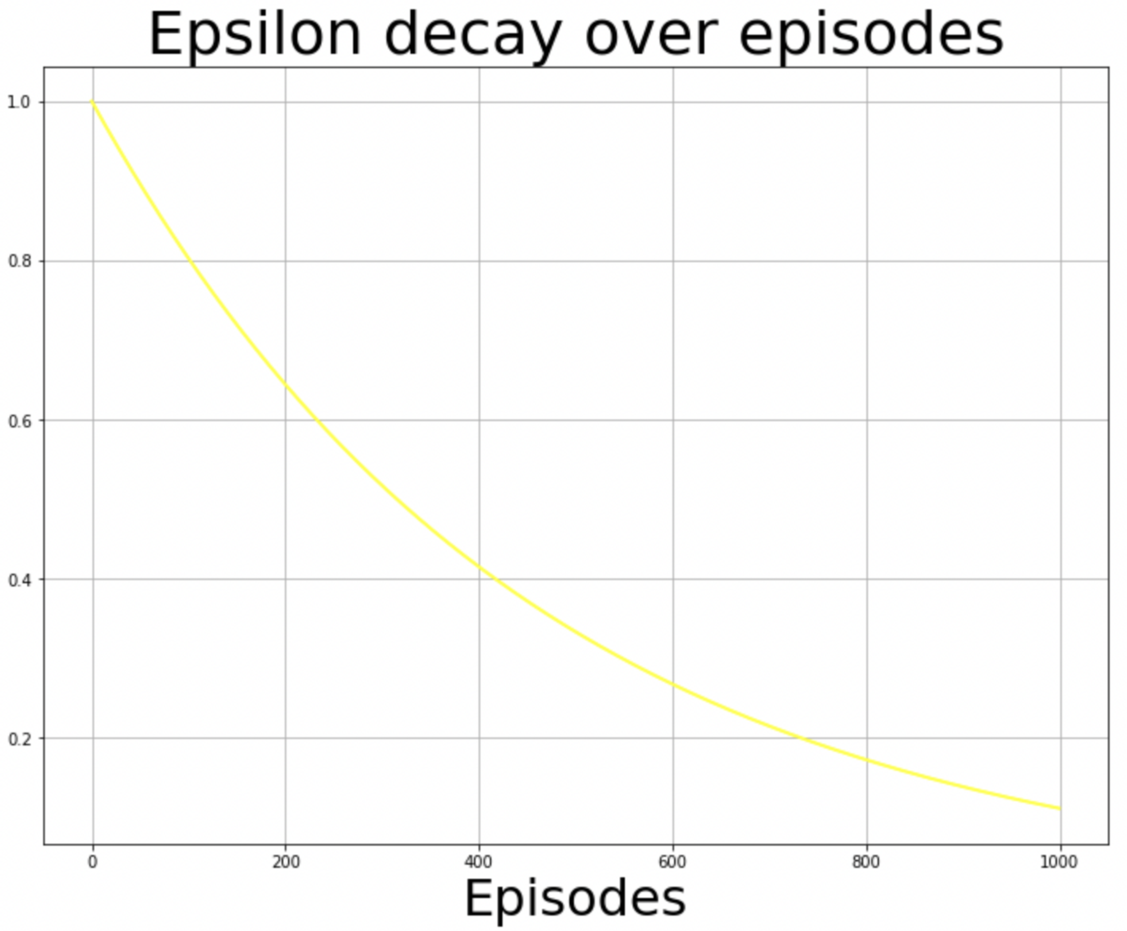


Fig 5 - Epsilon Decay over Episodes

Explanation -
We start with value of `ε = 1.` This value is updated as **epsilon = epsilon * decayRate.** Where Decay rate is 0.9978. Large epsilon values favour exploration and low epsilon values favour exploitation.