Experimenting Q-Learning Algorithm with I.I.D. Stock Model and Markov Stock Model

Group 10

November 23, 2018

Abstract

The group generated two sets of stock return rate data using different generation method - one with an I.I.D process and one with Markov process. We applied the Q-learning algorithm for each data set and discovered that the strategy learned based on each kind of data was different.

For I.I.D. data, Q-learning could not conclude a determined strategy: due to the fact that no correlation existed between adjacent data points of a single stock, the Q-learning could not predict the future behavior of the stock and therefore cannot effectively choose an action to make - the choice it made demonstrate randomness. Only if the I.I.D data set contained a stock that is obviously safe (for example, bond), the model would develop a strategy of choosing that stock.

For Markov data set, Q-learning could always find the best strategy.

1 Design Considerations and Assumptions

1.1 Modeling the Stock Environment

When modeling the stock environment, the team created four equity. For a given equity, the team generated its daily return rate satisfying either I.I.D. properties or Markov properties. The set of daily return rate of all equity was then fed into Q-learning algorithms as the observation of the environment.

When quantizing the state space, the team assumed any return rate $r \in [-0.05, 0.05]$ to be $r = 0$, $r \in [-1, -0.05)$ to be $r = -1$ and $r \in (0.05, 1]$ to be $r = 1$. This assumption was made due to the fact that an effect Q-learning process requires state space to be a finite space.

1.2 Modeling the Interaction with the Environment

In the physical world, stock buyers invest in equities by dividing his existing capital for each equity. Therefore, the action could be modeled as the proportions of capital for each equity where all proportions sum up to 1. However, to simplify the experiment, the team simplified the action by limiting the possible choice of proportion as 100% and 0%. This meant that, for each time step, the machine would invest 100% of the capital into one of the equities and invest 0% of the capital into other equity. This assumption was not just to simplify the situation but also due to the fact that an effect Q-learning process requires action space to be a finite space.

Secondly, the team assumed the investment cycle to be *oneday*. This means the buyer would not hold any equity for more than one day and he buys or sells the equities on a daily basis.

For the goal of the interaction, the team assumed the winning condition was to double the capital. In other words, the team was testing if the Q-learning can make correct decisions to double the initial capital.

1.3 Summary of Key Assumptions

- Any return rate $r \in [-0.05, 0.05]$ was assumed to be $r = 0$, $r \in [-1, -0.05)$ was assumed to be $r = -1$, and $r \in (0.05, 1]$ was assumed to be $r = 1$
- The return rate is modeled as daily return rate.
- The proportion of the capital investing can only be 100% or 0%, i.e buyers can only buy one equity at any time t.
- The investment cycle is assume to be *oneday*. The buyer would not hold any equity for more than one day and he buys or sells the equities in a daily basis.

Some of the assumptions above were not just made to simplify the situation but also due to the fact that an effect Q-learning process requires finite state space and action space. These assumptions also brought constraints to the following experiment. In the future, the team would continue to optimize the model to better simulate the real situation.

2 Experiments for I.I.D. Data

2.1 Data Generating Process

In I.I.D data set, the stock return rate at time t is independent of the stock return rate at time $t - 1$ for a given stock.

Therefore, when generating the data points of general equities, stocks, the team generated the data points that were uniformly distributed over the interval $[-1, 1]$. In other words, any value within the given interval was equally likely to be generated.

When generating the data for bonds, the team generated the return rate that was consistently $r = 0.05$.

When generating the data for the dummy equity indicating Not-investing, the team generated the return rate that was consistently $r = 0.00$.

2.2 Experiment I: Four I.I.D. Equities

In this experiment, the team generates I.I.D. equities as four general stocks. Every one of the stocks has a I.I.D source and therefore no correlation can be found between two of the data points.

The learning result met the team's expectation that the Q-learning algorithm's strategies demonstrated randomness and could not determined which stock was the most likely to gain profit.

The above result demonstrated that Q-Learning could not effectively determine an effective investment policy within 100 days of investment.

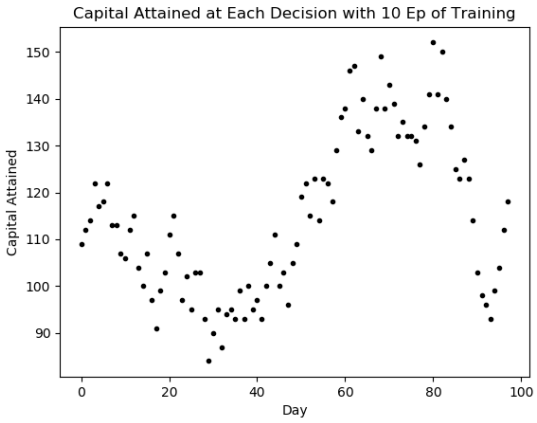

Figure 1: Four Stocks with I.I.D. Source

2.3 Experiment II: Three I.I.D. Stocks with One I.I.D. Dummy Equity

In this experiment, the team generated I.I.D. equities as one dummy equity(not-investing) and three general stocks. The result met the team's expectation that the Q-learning algorithm to indicate that consistently choosing dummy equity was the best strategy - because dummy equity was the most predictable equity and thus the safest one.

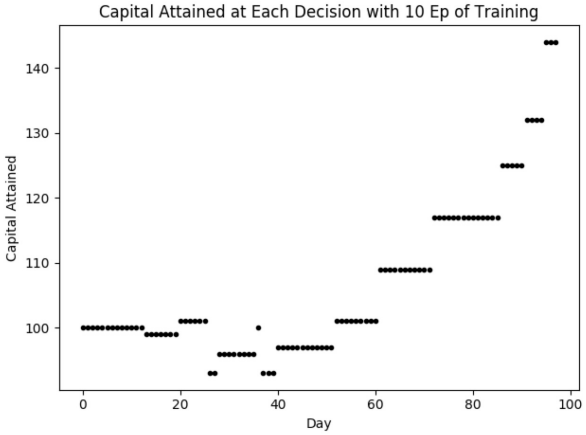

Figure 2: Four I.I.D. Equities with One of Them Indicating Not-investing

After the team introduced a dummy stock[Figure 2], with constant return rate 0% among i.i.d sources to represent the choice of not investing, the Q-learning table often computed a maximum expected future reward in choosing not to invest.

The Q-table of this trial of experiment was attached as Figure 3. By examining the Q-table, the team can see that the program had learned to reward the

action of choosing dummy equity $Equity0$. In other word, it had successfully developed a policy of choosing the most predictable equity, the dummy equity. This met the team's expectation.

	0	1	2	3
[0, 1, 1, -1]	0.315	-0.032	-4.000000e-02	-0.127
[0, -1, 1, -1]	0.293	-0.013	4.600000e-02	-0.149
[0, -1, 0, 0]	0.188	0.001	-9.000000e-03	-0.026
[0, -1, -1, -1]	0.102	-0.123	7.690000e-01	0.088
[0, 0, -1, -1]	0.279	-0.045	-7.100000e-02	-0.153
[0, 1, -1, -1]	0.333	-0.159	-1.380000e-01	-0.210
[0, 1, 1, 0]	0.239	0.069	1.800000e-02	0.066
[0, -1, 1, 1]	0.180	0.739	1.080000e-01	0.099
[0, 1, 0, -1]	0.227	0.008	-1.130000e-01	-0.116
[0, 1, -1, 1]	0.306	-0.279	1.750000e-01	0.230
[0, 1, -1, 0]	0.245	-0.176	-2.700000e-02	-0.024
[0, -1, 1, 0]	0.228	0.021	-1.040834e-17	-0.072
[0, -1, -1, 0]	0.240	-0.072	3.000000e-02	-0.074
[0, 1, 1, 1]	0.304	-0.113	1.280000e-01	0.047
[0, -1, -1, 1]	0.313	-0.400	-6.800000e-02	0.183
[0, -1, 0, -1]	0.217	0.003	7.600000e-02	-0.014
[0, 0, 1, 1]	0.272	-0.092	-5.200000e-02	-0.119
[0, 1, 0, 1]	0.251	0.054	1.500000e-02	-0.108
[0, -1, 0, 1]	0.246	-0.252	-9.000000e-03	-0.090
[0, 0, 1, -1]	0.259	0.117	1.700000e-02	-0.024
[0, 0, -1, 1]	0.265	0.024	-1.720000e-01	-0.044
[0, 0, 0, 0]	0.058	0.003	-6.000000e-03	0.000
[0, 1, 0, 0]	0.117	-0.020	4.300000e-02	-0.010
[0, 0, 0, 1]	0.023	-0.142	-8.000000e-02	0.496
[0, 0, 1, 0]	0.044	-0.002	-4.400000e-02	0.466
[0, 0, -1, 0]	0.137	-0.111	8.000000e-03	-0.037
[0, 0, 0, -1]	0.127	-0.061	1.400000e-02	-0.179

Figure 3: Four I.I.D. Equities with One of Them Indicating Not-investing

2.4 Experiment III: Three I.I.D. Stocks with One of I.I.D. Bond

In this experiment, the team generated I.I.D. equities as one bond and three general stocks. The result met the team's expectation that the Q-learning algorithm consistently chose bond was the best strategy - because dummy equity was the most predictable equity and thus the safest one.

As shown in above figure, if the team introduced a fixed asset, or bond, the Q-learning developed a policy that it would often choose to invest in it, rather than the stocks with i.i.d return rates. A further demonstration can be seen in the Q-table [Figure 5], in which the bond equity was always given the highest action reward.

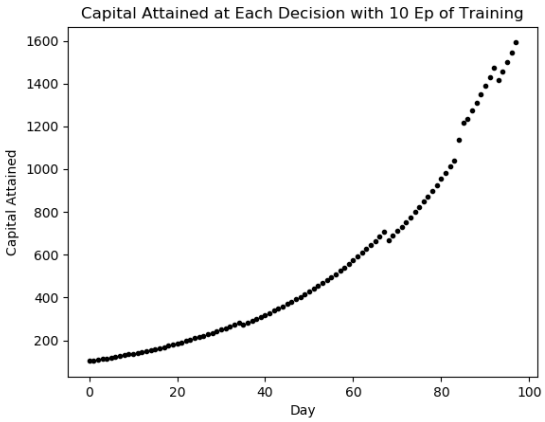

Figure 4: Four I.I.D. Equities with One of Them Indicating Bond Equity

	0	1	2	3
[1, 1, 1, -1]	1087.949	498.872	483.815	452.315
[1, -1, -1, -1]	1107.361	473.717	419.884	421.754
[1, -1, -1, 1]	1091.790	422.314	423.848	477.435
[1, 1, -1, 1]	1098.669	435.154	451.464	475.482
[1, 1, 1, 1]	1104.692	494.008	524.850	493.510
[1, 1, -1, -1]	1100.904	440.044	470.403	418.929
[1, -1, 1, -1]	1088.332	430.772	477.691	394.396
[1, 0, 1, 1]	994.786	84.976	164.860	112.659
[1, -1, 1, 1]	552.849	430.796	496.484	887.520
[1, 1, -1, 0]	953.908	87.099	95.245	159.149
[1, 1, 1, 0]	194.808	833.225	133.860	126.268
[1, 1, 0, 1]	132.379	832.683	115.644	137.589
[1, 1, 0, -1]	1014.906	171.682	148.190	92.419
[1, -1, 1, 0]	995.654	142.662	49.078	197.679
[1, -1, 0, -1]	1000.820	147.523	74.024	106.181
[1, -1, -1, 0]	1012.396	138.800	92.147	100.138
[1, -1, 0, 1]	995.440	98.560	131.293	131.781
[1, 0, -1, 1]	991.051	95.098	98.156	124.418
[1, 0, 0, 1]	749.069	12.462	56.661	7.378
[1, 0, -1, -1]	157.569	37.470	800.171	74.958
[1, 0, -1, 0]	717.858	40.974	21.735	16.133
[1, 1, 0, 0]	726.423	17.176	32.383	26.834
[1, 0, 1, -1]	989.865	95.588	114.020	117.487
[1, -1, 0, 0]	602.623	32.468	19.733	21.610
[1, 0, 1, 0]	650.564	32.960	42.602	39.207
[1, 0, 0, -1]	761.231	16.051	20.496	27.124
[1, 0, 0, 0]	104.962	0.000	6.999	4.961

Figure 5: Q-table of Four I.I.D. Equities with One of Them Indicating Bond Equity

2.5 Experiment IV: Two I.I.D. Stocks with One I.I.D. Bond and One I.I.D Dummy Equity

In this experiment, the team generated I.I.D. equities as one dummy equity, one bond, and two general stocks. As expected, due to the predictability, the algorithm chose bond and not-invest most of the time. Comparing with Experiment III, having the option of not-invest reduces the profitability.

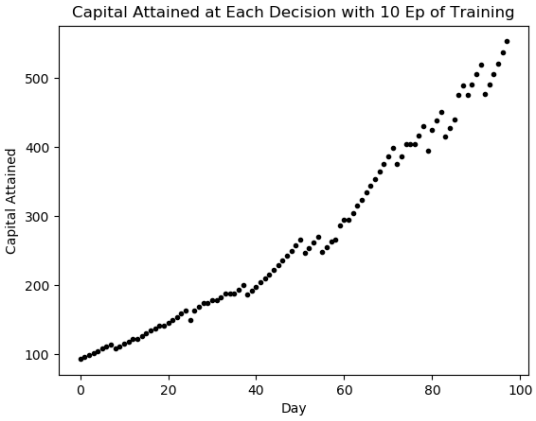

Figure 6: Four I.I.D. Equities with One of Them Indicating Not-investing and One of Them Indicating Bond Equity

When both dummy equity and bond were introduced, the Q-learning developed a strategy similar with Experiment III, invest in bond, while also adding action of not-investing sometimes.

3 Experiments for Markov Data

3.1 Data Generating Process

In Markov data set, the stock return rate at time t is dependent on the stock return rate at time $t-1$ for a given stock.

Therefore, when generating the data points of equities, the team first made an assumption for the potential return rates of the given stock, then the team made an assumption for the 1-step transition matrix.

For this sets of the experiment, the team generated the data for three kinds of equity: bond, stocks, and dummy equity indicating not-investing.

For a low reward but more predicable equity - bond, the team assume the possible return rates to be

$$\begin{bmatrix} -0.05 & 0.00 & 0.05 \end{bmatrix}$$

with a transition matrix

$$\begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$$

The team denoted this kind of equity as $Stock0$.

For a larger reward but less predictable equity, the team assume the possible return rates to be

$$\begin{bmatrix} -0.1 & 0.00 & 1 \end{bmatrix}$$

with a transition matrix

$$\begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

The team denoted this kind of equity as $Stock1$.

For a big reward but even less predictable equity, the team assume the possible return rates to be

$$\begin{bmatrix} -0.25 & 0.00 & 0.25 \end{bmatrix}$$

with a transition matrix

$$\begin{bmatrix} 0.2 & 0.4 & 0.4 \\ 0.4 & 0.2 & 0.4 \\ 0.4 & 0.4 & 0.2 \end{bmatrix}$$

The team denoted this kind of equity as $Stock2$.

For a big reward but even less predictable equity, the team assume the possible return rates to be

$$\begin{bmatrix} -0.25 & 0.00 & 0.25 \end{bmatrix}$$

with a transition matrix

$$\begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.9 & 0.05 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$$

The team denoted this kind of equity as $Stock3$.

For the dummy equity indicating not-investing, the team generated the return rate that was consistently $r = 0.00$. The team denoted this kind of equity as $Dummy$.

3.2 Experiment I: Four Stocks of Markov Source

Consider now the return rates being modeled by a Markovian process of memory 1. As expected, this yielded better results than the i.i.d scenario.

In this experiment, the team generated four Markov source stocks: $Stock0$, $Stock1$, $Stock2$, $Stock3$. Therefore in this case the algorithm would not have the option of not-invest. Obviously, the $Stock3$ with largest possible return rate and highest predictability was the optimal solution. Let us see what policy Q-learning developed.

As expected, the program developed a policy of investing the stock with highest probability of higher profit. We can examine the Q-table and see if the program was making the correct decision in the process.

A partial Q-table was shown in Figure 6. We can see that near the end of the process, the program has learn the correct policy that it gave lowest reward to the state (x_3 = -0.25, a_3 = "invest all capital in $Stock3$")

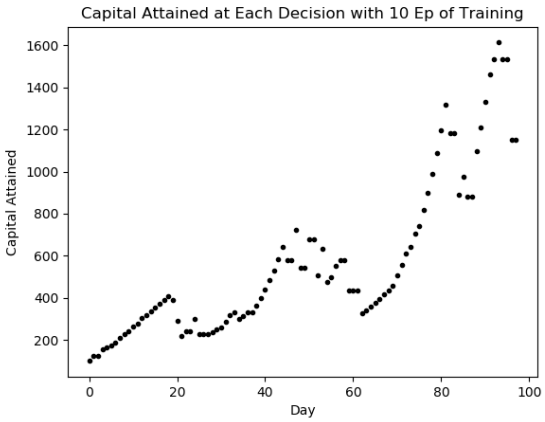

Figure 7: Four Markov Source Stocks

	0	1	2	3
[-0.05, 0.0, 0.25, 0.0]	-7.537	15.794	13.367	190.340
[0.0, 0.0, 0.0, 0.0]	5.622	10.490	7.051	158.386
[0.0, 0.0, -0.25, 0.0]	13.329	17.392	190.895	12.612
[0.0, 0.0, 0.25, 0.0]	2.799	4.877	9.132	154.476
[-0.05, 0.0, 0.0, 0.0]	-8.861	206.272	17.485	21.983
[-0.05, -0.1, 0.0, 0.0]	-9.550	-7.094	18.028	217.471
[-0.05, 0.0, -0.25, 0.0]	8.221	23.909	249.127	8.964
[-0.05, 0.1, 0.25, 0.0]	22.320	542.160	24.494	48.332
...	...	...	...	...
[-0.05, -0.1, 0.0, 0.25]	15.238	57.827	59.525	929.165
[0.05, -0.1, -0.25, 0.25]	75.167	22.399	60.975	653.630
[0.05, -0.1, 0.25, 0.25]	69.382	26.227	12.631	657.555
[0.05, 0.1, 0.25, -0.25]	52.800	639.063	52.346	16.549
[0.0, 0.1, -0.25, 0.0]	21.855	407.432	33.583	19.487
[0.0, -0.1, 0.0, 0.0]	9.583	-8.857	-6.164	156.187
[-0.05, -0.1, 0.0, -0.25]	-5.592	9.210	228.615	-28.111
[0.05, -0.1, 0.25, -0.25]	614.735	14.768	29.634	14.570

Figure 8: Q-table with Four Markov Source Equities

3.3 Experiment II: Three Markov Source Stocks With One Dummy Equity

In this experiment, the team generated three Markov source stocks With one Dummy equity. As expected, this yeilds better results than the i.i.d process. The capital earned in this scenario was the highest amount among all the tests.

The capital earned in this trail was higher than the previous case. The team believed that the existence of Dummy equity in this experiment helped the algorithm to preserve capital when the risk of investing is high, which helped to achieve higher profit.

4 Conclusion and Analysis

According to the experiment result, the team concluded that Q-learning worked better for Markov data and less desirable for I.I.D. data in general. For a

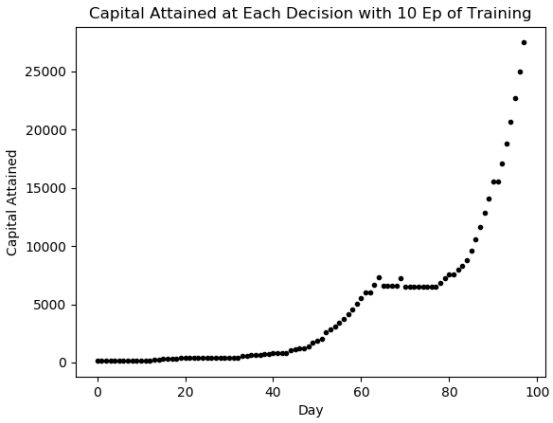

Figure 9: Four Markov Source Stocks

Markov data set, the Q-learning algorithm could always learn the best strategy. However, for a I.I.D. source, the effect of Q-learning algorithm depends on the combination of the equities - for most of the time, it tended to choose the best or the safest option, either not-invest or bond.

Note that, some of the experiment assumptions or implementation decision might have limited the discovery or caused unexpected result for the experiments.

Recall that any return rate $r \in [-0.05, 0.05]$ was assumed to be $r = 0$, $r \in [-1, -0.05)$ was assumed to be $r = -1$, and $r \in (0.05, 1]$ was assumed to be $r = 1$. This threshold of discretizing the return rates might not best simulate the reality. Under this threshold, an equity with $r = 0.06$ and an equity with $r = 1$ would be considered as two equities with the same profitability at a given time t. The team might need to consider the effectiveness of the threshold.

Another decision was that the return rate was modeled as a daily return rate. In reality, it is obvious that the stock return rate varies more within one day than within one hour. In other words, the stock return rate generated in the current experiments was more similar to an I.I.D random variable. If hourly data was observed and in a high-frequency trading scenario, Q-learning was expected to work better.

In terms of action, the team assumed that buyers can only buy one equity and on a daily basis. This reduces the possible combination of actions and thus reduce the power of Q-learning.

Moving forward, the team will introduce more equity combination in order to find the best portfolio strategy. The team will also continue to optimize the model so that it better simulates the real situation, which help the team can further understand how Q-learning works for I.I.D. and Markov source data.