پاسخ تکلیف سوم

"یادگیری تعاملی"

گروه دوم

نام استاد:

دکتر نیلی

تهیه کننده :

عرفان میرزایی ۸۱۰۱۹۹۲۸۹

تکلیف سوم- یادگیری تعاملی
عرفان میرزایی
۸۱۰۱۹۹۲۸۹

پاسخ پرسش ۱:

۱. مدلسازی:

در این مسئله باید ابتدا MDP مورد نظر برای سوال را تعریف کنیم. ما در این مدلسازی حالت را در هر مرحله یک چهارتایی مرتب شامل قیمت هر یک از سهام ها به ترتیب و در آخر میزان دارایی عامل در نظر گرفتیم. که ارزش سهام شرکت ها عددی بین ۱ تا ۱۰ است (میزان ارزش سهام شرکت به دلار تقسیم بر ۵) و دارایی عامل عددی از بین ۱ تا ۲۰ (یعنی از ۵ تا ۱۰۰ دلار) هم چنین عملی که در هر مرحله می‌توانیم انجام دهیم را به صورت یک سه تایی مرتب در نظر گرفتیم که هر کدام از مولفه های آن می‌تواند صفر یا یک باشد. عملی که همه‌ی مولفه‌های آن یک باشد را خارج از مجموعه‌ی اعمال در نظر می‌گیریم.

مدل ارائه شده ویژگی مارکوف را دارد به این معنا که وقتی در یک حالت قرار داریم با فرض انجام یک عمل مشخص، حالت بعدی و پاداش دریافتی کاملا مستقل از حالت ها و اعمال ما در زمان های گذشته است.

حال باید احتمال گذار هر یک از حالت ها به حالت دیگری را تعریف کنیم که این احتمال ها با توجه به جدول های داده شده در صورت سوال و هم چنین ارزش اولیه و ثانویه سهام که از حالت ها بدست می‌آید با فرض مستقل بودن تغییر قیمت شرکت ها از یکدیگر به سادگی قابل محاسبه است. هم چنین با توجه به مدل پیاده‌سازی شده برای گذار از یک حالت به حالت دیگر تنها یک پاداش وجود دارد که مقدار آن از تفاوت دارایی های عامل در دو حالت بدست می‌آید.

نکات و فرضیات :

۱. در این مسئله در نظر گرفته شده است که عامل در ابتدای هر مرحله‌ی زمانی سهام را خریده و در پایان آن روز سهام را در هر صورتی می‌فروشد و دارایی عامل به روز رسانی می‌شود.
۲. با توجه به صورت مسئله حالت های نهایی را زمانی در نظر گرفته ایم که اگر دارایی عامل به ۱۰۰ یا ۵ دلار برسد مسئله خاتمه می‌یابد و در مدل ما در این صورت به ترتیب عامل پاداشی برابر +۱۰۰ و یا -۱۰۰ دریافت می‌کند که معادل +۵۰۰ و -۵۰۰ دلار می‌باشد.
۳. در این مسئله با توجه به آنکه احتمال تغییر ارزش سهام با توجه به مقدار آن می‌تواند متفاوت باشد(در حداکثر و حداقل ارزش) مجبوریم که این مقادیر را جز حالت های عامل در نظر بگیریم.
۴. در این مدل سازی در نظر گرفته شده است که اگر عامل عمل خرید سهامی را انتخاب کرد که بیشتر از دارایی آن است ارزش شرکت ها با توجه به احتمالات تغییر کرده ولی دارایی عامل ثابت می‌ماند.

۲. پیاده سازی :

پس از مدل سازی مسئله در قالب MDP باید آن را پیاده سازی کنیم. برای حل یک MDP یعنی بدست آوردن سیاست بهینه مجبور به استفاده از برنامه نویسی پویا هستیم که در این سوال از روش Policy iteration استفاده می‌کنیم. به این منظور از شبه کد زیر که مربوط به کتاب ساتون و بارتو است استفاده شده است.

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))\left[r + \gamma V(s')\right]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)\left[r + \gamma V(s')\right]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow false$
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

شکل ۱- شبه کد Policy Iteration

که به این منظور کلاس محیط StockMarketEnvironment را پیاده سازی می‌کنیم. این کلاس از EnvironmentBase مشتق شده است و ویژگی های آن را ارث بری کرده است. مجموعه حالت ها و اعمال به صورت گفته شده در بخش قبلی درنظر گرفته شده است که بدین منظور از MultiDiscrete کتابخانه‌ی gym استفاده کردیم. مقدار نرخ تخفیف و هم چنین احتمال گذارها را به عنوان ورودی دریافت می‌کنیم.

شبه کد پیاده سازی شده را در متد Policy Iteration پیاده سازی کرده ایم که این متد یک تابع ارزش اولیه و یک سیاست قطعی را به عنوان ورودی دریافت می‌کند و سیاست و تابع ارزش بهینه را به عنوان خروجی باز می‌گرداند. هم چنین مقدار تتا را نیز به عنوان ورودی دریافت می‌کند که دقت ارزیابی ارزش ها را مشخص می‌کند.

با استفاده از متد get_all_states, get_available_actions تمام حالت ها و عمل های ممکن را در خروجی باز می‌گردانیم. برای پیاده سازی این شبه کد از یک حلقه‌ی While بزرگ استفاده کردیم که تا وقتی سیاست پایدار نشده یعنی به ازای انجام مرحله‌ی بهبود سیاست، سیاست مربوط برخی از حالت ها تغییر می‌کند این کار را ادامه می‌دهیم. بدین صورت که در هر مرتبه ابتدا ارزش ها را به روزرسانی کرده تا زمانی که بیشترین اختلاف میان دو ارزش یک حالت به کمتر از تتا برسد. برای به روز رسانی ارزش یک حالت می‌بایست که احتمال گذار از یک حالت به حالت بعدی و هم چنین پاداش دریافتی در آن حالت را داشته باشیم. که با توجه به آنکه مدل ما دارای ۲۰۰۰۰ حالت است بدین صورت عمل کرده ایم که تنها این مقادیر را برای حالت های همسایه یک حالت محاسبه کردیم و این کار را در متد get_dynamics پیاده سازی کرده‌ایم. بدین صورت که در هر حالت-عمل حالت های همسایه و احتمال رفتن به این حالت ها و هم چنین پاداش گذار به این حالت را به عنوان خروجی باز می‌گردانیم. که برای تولید حالت های مجاور، از شروط استفاده می‌کنیم.

هم چنین پس از بدست آوردن این احتمالات و پاداش ها باید مقدار ارزش هر حالت را به روز رسانی کنیم که این کار را به کمک متد update_values انجام می‌دهیم که به ازای همه‌ی حالت های مجاور مقدار احتمال را در مجموع پاداش و ارزش کاهش یافته حالت مجاور ضرب می‌کند و مقادیر آن ها را با یکدیگر جمع می‌کند.

پس از پیاده سازی متد Policy Iteration در نوت بوک Q1.ipynb که مربوط به این سوال است ابتدا کتابخانه های لازم را وارد می‌کنیم. سپس احتمال های تغییر ارزش سهام شرکت ها و هم چنین عمل ها را به عنوان ورودی های لازم برای پاس دادن به محیط تعریف می‌کنیم. هم چنین یک سیاست قطعی که در هر حالت به تصادف یک عمل را انتخاب می کند را به همراه تابع ارزش صفر برای هر حالت به عنوان ورودی به الگوریتم Policy Iteration می دهیم.

۳. نتایج :

برای با خبر شدن از میزان پیشرفت الگوریتم در هر بار اجرای حلقه‌ی خارجی الگوریتم به ازای هر بار که الگوریتم Value Evaluation اجرا می‌گردد میزان دلتا که با تتا مقایسه می‌شود را چاپ می‌کنیم که همان طور مشاهده می‌شود در هر بار تلاش این مقدار کاهش می‌یابد و هم‌چنین برای هر بار اجرای حلقه خارجی نیز این مقادیر

کوچکتر از مقادیر گذشته است که این(کوچک تر شدن تغییرات ارزش ها) می‌تواند نشانه ای از نزدیک به همگرا شدن الگوریتم باشد.

هم چنین در مرحله Policy Improvement با استفاده از یک شمارنده تعداد حالت هایی که سیاست آن عوض شده اند را حساب کرده و چاپ می‌کنیم. که همان طور که انتظار داشتیم این مقدار برای اولین مرحله بدلیل تصادفی بودن مقدار بسیار بالایی دارد (۱۸۰۰۰) و رفته رفته و در طی پنج گام این تغییرات به صفر می‌رسد که این امر نشان دهنده‌ی این است که با توجه به ارزش های حالات که با دقت تتا بدست آمده است توانسته ایم که سیاست بهینه را بدست بیاوریم.

تتا در تمامی مراحل به مقدار ۰,۰۲ محاسبه شده است که این مقدار با توجه به اینکه ارزش پاداش ها مقداری بین +۱۰۰ و +۱ و ۰ و -۱ و -۱۰۰ است مقدار کوچکی به محسوب می‌شود.

با توجه به اینکه در این مسئله احتمال گذار بین حالت های مختلف را داشتیم می‌توانیم امید ریاضی هر عمل را محاسبه کنیم با این فرض که در هر حالت ما قادر به انجام تمامی اعمال هستیم. که می‌دانیم این فرض با توجه به اینکه ارزش شرکت ها متغییر با زمان است و هم‌چنین دارایی عامل نیز تغییر می‌کند در همه‌ی شرایط برقرار نیست.

$$\boldsymbol{E}[Reward(0,0,0) = 0$$

$$\boldsymbol{E}[Reward(1,0,0) = 0.4 * (+5) + 0.3 * (0) + 0.3 * (-5) = +0.5$$

$$\boldsymbol{E}[Reward(0,1,0) = 0.1 * (+5) + 0.8 * (0) + 0.1 * (-5) = 0$$

$$\boldsymbol{E}[Reward(0,0,1) = 0.2 * (+5) + 0.1 * (0) + 0.7 * (-5) = -2.5$$

$$\boldsymbol{E}[Reward(1,1,0) = 0.04 * (+10) + 0.35 * (+5) + 0.31 * (0) + 0.27 * (-5)$$
$$+0.03 * (-10) = +0.5$$

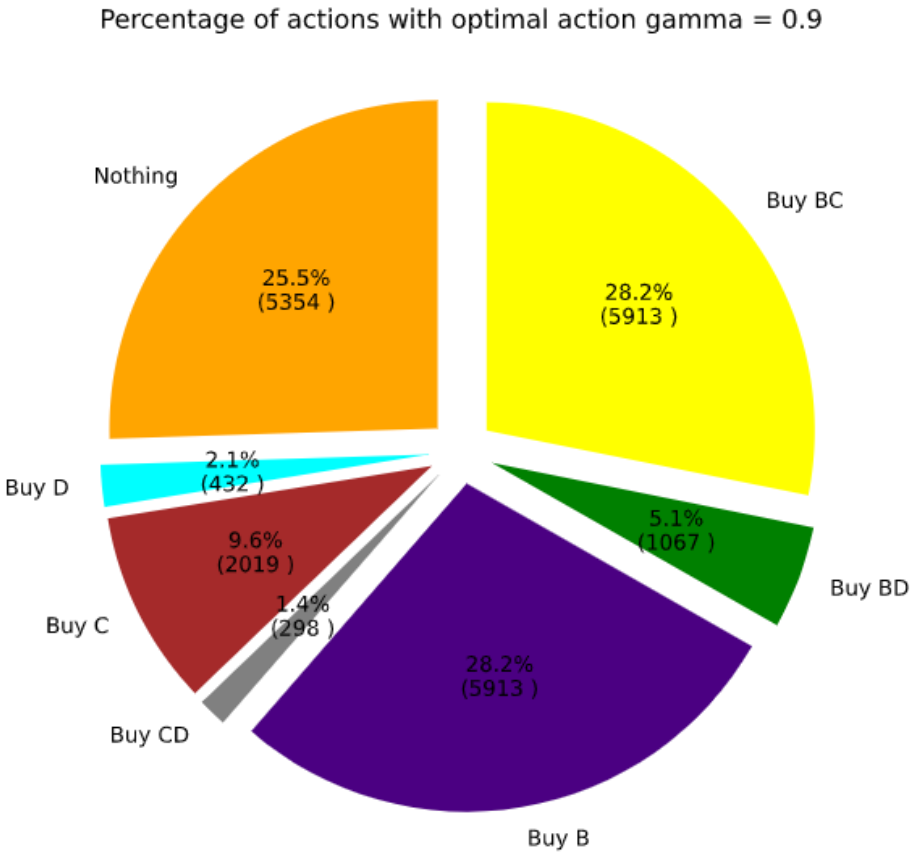
$$\boldsymbol{E}[Reward(1,0,1) = 0.08 * (+10) + 0.1 * (+5) + 0.37 * (0) + 0.24 * (-5)$$
$$+0.21 * (-10) = -2$$

$$\boldsymbol{E}[Reward(0,1,1) = 0.02 * (+10) + 0.17 * (+5) + 0.16 * (0) + 0.57 * (-5)$$
$$+0.08 * (-10) = \text{-}2.6$$

مولفه‌ی اول برای خرید شرکت B است و مولفه ی برای خرید سهام شرکت C و هم چنین D می باشد.

که با توجه به این مقادیر در هر مرحله در صورتی که بتواند تمام عملیات ها را انجام دهد می‌تواند بین حالت خرید سهام شرکت B و یا خرید سهام شرکت B , C انتخاب کند.

حال برای ارزیابی سیاست بهینه‌ی بدست آمده میزان انتخاب عمل های متفاوت را در یک نمودار دایره‌ای چاپ می‌کنیم.

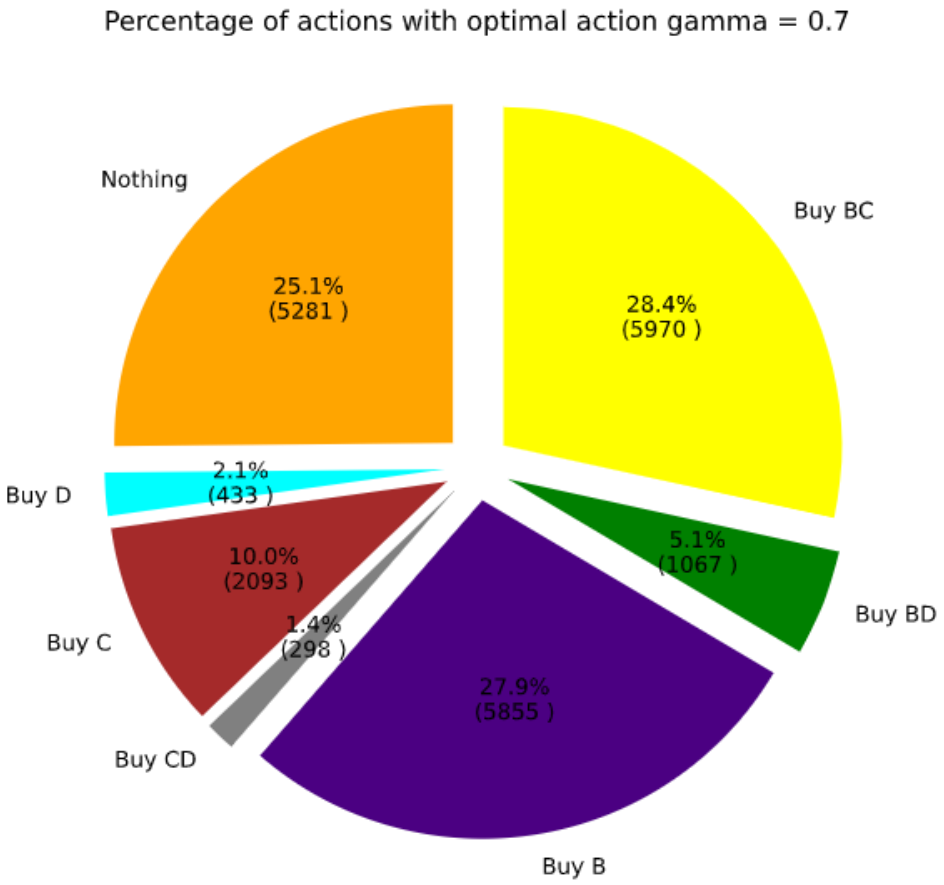


که همان طور که انتظار میرفت عمل خرید سهام B به همراه خرید سهام B , C به یک اندازه و در بالاترین میزان قرار دارند. که جمع آن ها تقریبا بیش از نیمی از حالت ها را به خود اختصاص می‌دهند که با یک محاسبه‌ی سر انگشتی می‌توان در نظر گرفت وقتی که دارایی ما ۱۱ تا ۲۰ باشد که نیمی از حالت ها را دربر می‌گیرد سهام شرکت ها هر چه باشد ما می‌توانیم تمام اعمال را انجام بدهیم و برای بقیه دارایی های از بین ۰ تا ۱۰ نیز عامل در برخی از شرایط می‌تواند عمل بهینه را انجام بدهد که به همین دلیل جمع آن ها بیشتر از ۵۰ درصد است.

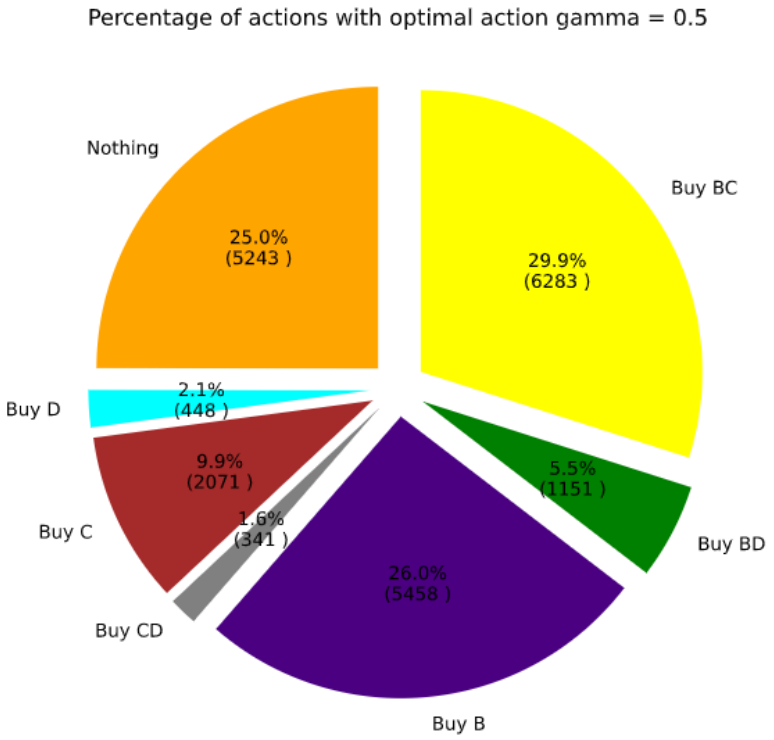
در جایگاه بعدی عمل سهامی نخریدن قرار دارد که مربوط به حالت هایی می‌شود که دارایی های فرد کمتر از قیمت سهام هر یک از شرکت هاست و فرد نمی‌تواند سهامی خریداری کند. هر چند که در این شرایط عامل می‌تواند هر عملی را انجام بدهد زیرا نتیجه همواره یکسان خواهد بود.

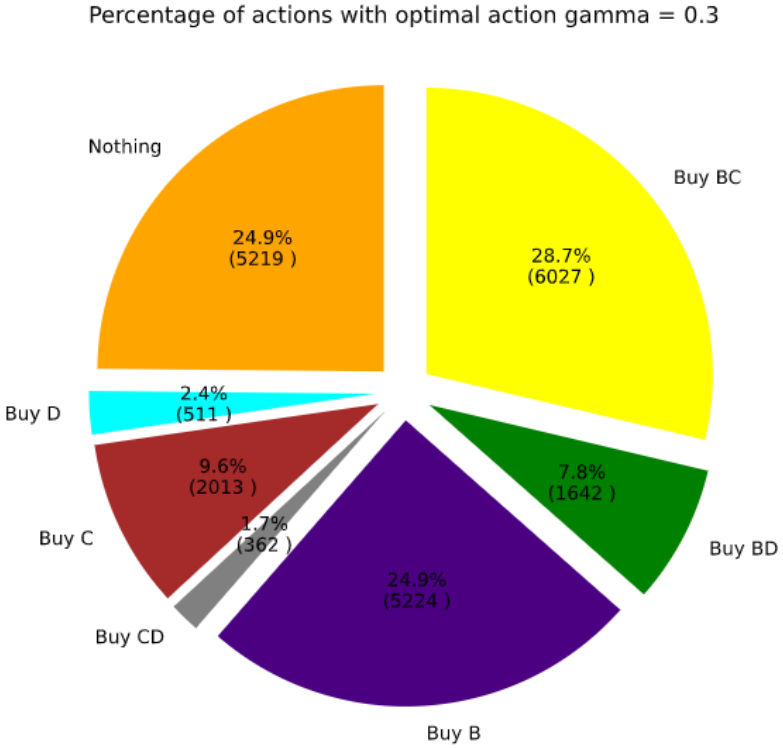
هم چنین دیده می‌شود که ترتیب سایر اعمال نیز دقیقا به همان ترتیبی است که امید ریاضی آن ها را حساب کردیم و مربوط به موقعیت هایی است که به دلیل کم بودن دارایی از ارزش سهام شرکت ها می‌توانسته ایم هر عملی را انجام بدهیم.

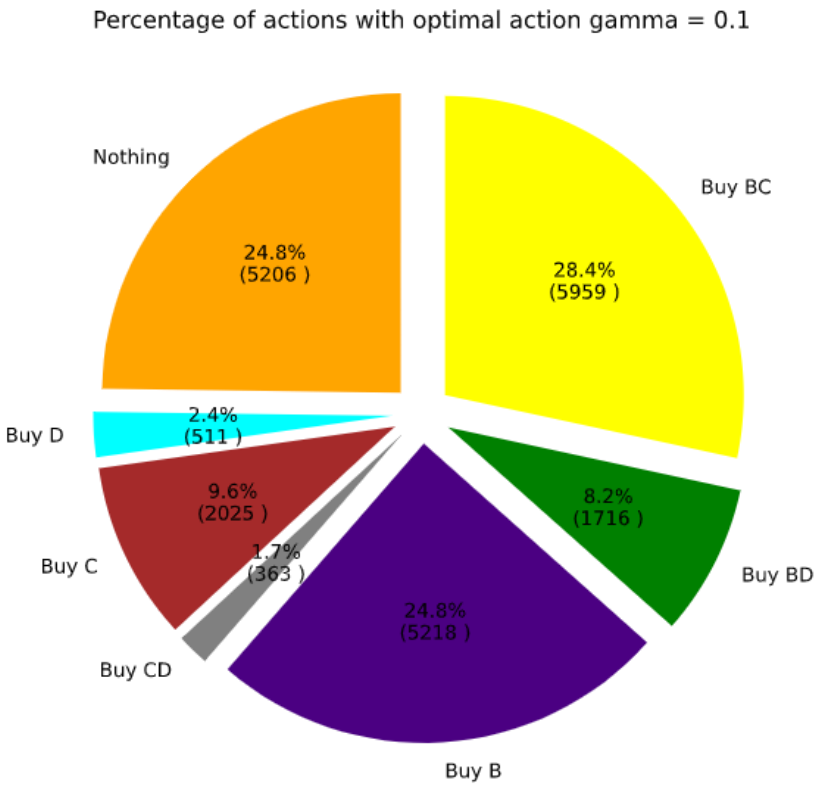
۴. مقایسه‌ی نتایج :



تکلیف سوم- یادگیری تعاملی
عرفان میرزایی
۸۱۰۱۹۹۲۸۹







Gamma	۰,۹	۰.۷	۰.۵	۰.۳	۰.۱
B	۲۸.۲	۲۷.۹	۲۶	۲۴.۹	۲۴.۸
B, C	۲۸.۲	۲۸.۴	۲۹.۹	۲۸.۷	۲۸.۴
No	۲۵.۲	۲۵.۱	۲۵	۲۴.۹	۲۴.۸
C	۹.۶	۱۰	۹.۹	۹.۶	۹.۶
B, D	۵.۴	۵.۱	۵.۵	۷.۸	۸.۲
D	۲.۱	۲.۱	۲.۱	۲.۴	۲.۴
C, D	۱.۴	۱.۴	۱.۶	۱.۷	۱.۷

با توجه به اعداد ثبت شده در جدول و هم چنین نمودار های دایره‌ای می‌توان مشاهده کرد که عملکرد عامل ها در تمامی مراحل نزدیک به یکدیگر است اما با کاهش میزان نرخ بهره و کوته‌بین شدن عامل، مقداری از انجام عمل بهینه دست می‌کشد که این اتفاق به دلیل کم شدن اثر پاداش بزرگی است که با رسیدن به دارایی ۱۰۰ می‌رسیم صورت می‌گیرد.

پاسخ پرسش ۲ :

۱. فرضیات :

۱. در این مدلسازی فرض شده است که شرکت توانایی خرید هر مقداری از محصولات اولیه را در اول ماه دارد و محدودیتی وجود ندارد. با فرض اینکه در محیط شبیه سازی برای یادگیری این مسئله مشکلی ندارد و پس از اتمام یادگیری و در دنیای واقعی عامل یاد گرفته است که اگر ماده‌ی اولیه ای را بخرد ولی تاریخ انقضای آن تمام شود ضرر می‌کند.

۲. فرض کرده‌ایم که سهولت در تهیه‌ی ماده اولیه و یا تاثیر تحریم بودن یا نبودن یک ماده‌ی اولیه بر روی قیمت آن منعکس می‌شود و در صورت نایاب بودن ماده‌ی اولیه‌ای می‌توان قیمت آن را به صورت حدی بی‌نهایت درنظر گرفت.

۳. فرض کرده‌ایم که مواد اولیه‌ی لازم برای هر محصول ثابت است و با زمان تغییر نمی‌کند و با توجه به پاداشی که عامل در هر مرحله دریافت می‌کند خود می‌تواند به درستی یاد بگیرد که برای هر محصول چه میزان و چه ماده‌ی اولیه‌ای لازم است.

۴. در این مدلسازی فرض کرده‌ایم که سلیقه‌ی مشتریان در ماه های مختلف سال متفاوت است اما به طور کلی، در سال های متفاوت یکسان است.

۵. با توجه به فرض گفته شده در صورت سوال که خریدار می‌تواند برای خرید محصولات به سایت شرکت مراجعه کند، فرض کرده‌ایم که خریدار پس از ورود به سایت محصولات مورد انتظار خود را وارد می‌کند و در صورت آنکه آن محصولات در کارخانه موجود باشد به فروش میرسد و در همان لحظه به مشتری تحویل می‌شود و پول آن بلافاصله دریافت می‌شود و همچنین درصورتی که آن جنس موجود نباشد به مشتری اطلاع داده می شود و عامل نیز پاداش منفی ای دریافت می‌کند.

۶. فرض کرده‌ایم که عامل برای یک سال برنامه ریزی شده است و در انتهای سال شروع مجدد می‌شود. در انتهای هر سال انبار گردانی مواد اولیه اتفاق می‌افتد و مواد اولیه کارخانه را صفر می‌کنیم و هم چنین در پایان هر ماه انبار گردانی محصولات را انجام می‌دهیم و تمام محصولات را صفر می‌کنیم.

۷. در صورتی که عامل دستور تولید تعداد محصولی را بدهد که مواد اولیه‌ی آن در انبار نباشد، تا جایی که می‌توان از مواد اولیه استفاده می‌کنیم تا حدی که نشود یک محصول دیگر را آماده کرد.

۸. در صورتی که عامل دستور خرید محصولات غذایی را بدهد اما در روز اول ماه نباشیم هیچ تغییری در مواد اولیه رخ نمی‌دهد.

۹. فرض می‌کنیم در هر گام زمانی می‌توانیم تنها یک فروش از سایت انجام بدهیم.

۲. مجموعه حالت ها:

برای مدلسازی این مسئله با توجه به اینکه باید علاوه بر خرید مواد اولیه که در اول هر ماه اتفاق می‌افتد، تبدیل مواد اولیه به محصولات را نیز مدیریت کنیم و هم چنین فروش محصولات بر روی حالت عامل نیز تاثیر گذارند پس در نهایت گام زمانی را برابر یک دقیقه در نظر می‌گیریم.

یک ماتریس ۵۰۰ سطری در نظر می‌گیریم که در ستون اول مقدار کالای اولیه موجود در انبار است (که برای هر این اعداد مقیاسی در نظر گرفته شده است. برای مثال به گرم گزارش می‌کنیم) و در ستون دوم مدت زمان باقی مانده تا منقضی شدن آن(چند دقیقه) را نشان می‌دهد و ستون سوم آن قیمت آن ماده‌ی اولیه را بیان می‌کند.

ستون اول و دوم این ماتریس برای هر سطر یک لیست شامل ۱۲ عدد است که میزان آن ماده اولیه و تاریخ انقضای هر ماده اولیه در آن روز را با توجه به ماه خریداری شده‌ی آن نشان می‌دهد.

هم چنین یک ماتریس دیگری با ۱۰۰ سطر در نظر می‌گیریم که در ستون اول تعداد محصول موجود در انبار کارخانه و در ستون دوم آن مدت زمان باقی مانده تا منقضی شدن آن(چند دقیقه) را نشان می‌دهد و ستون سوم آن قیمت آن محصول را بیان می‌کند.

ستون اول و دوم این ماتریس برای هر سطر یک لیست شامل ۱۲ عدد است که تعداد آن محصول و تاریخ انقضای هر محصول را با توجه به دقیقه‌ی تولید آن نشان می‌دهد.

حالت محیط ما در این مسئله از متشکل از ماتریس هایی است که در بالا ذکر شده به اضافه‌ی یک عدد ثابت که نشان‌دهنده‌ی مقدار دقیقه گذشته از ابتدای سال است.

۵. مجموعه اعمال:

این عامل در هر حالتی که قرار دارد (یعنی در هر دقیقه) می‌تواند تعداد تولید از هر محصول را مشخص کند که یک بردار ۱۰۰ تایی است و هم چنین در ادامه با یک بردار ۵۰۰ تایی میزان خرید هر ماده‌ی اولیه را(به گرم) مشخص کند. در واقع هر عمل برداری شامل ۶۰۰ مولفه‌ای است.

۶. پاداش:

در مدلسازی مسئله ها به روش یادگیری تقویتی ما با استفاده از سیگنال پاداش برای عامل مشخص می کنیم که هدف "چیست" اما در مورد "چگونگی" رسیدن به هدف اطلاعی به عامل نمی‌دهیم. در این مثال نیز مانند سایر مسائل یادگیری تقویتی پاداش به صورت یک اسکالر متعلق به اعداد حقیقی در نظر گرفته شده که میزان سودی است که در ماه گذشته کسب کرده‌ایم. هدف عامل یادگیری تقویتی این است که امید ریاضی میزان پاداش های دریافتی را در طول زمان بیشینه کند. میزان سود که در واقع به عنوان پاداش به عامل داده می‌شود با انجام محاسبات پیچیده‌ای که توسط محیط(خارج از عامل) مشخص می‌شود؛ بدین صورت که مقدار فروش محصولات شرکت را در دقیقه اخیر حساب کرده و سپس هزینه‌ی تولید این محصولات شامل مواد اولیه خریداری شده را با علامت منفی در نظر می‌گیریم و با آن جمع می‌کنیم. هم‌چنین میزان سودی که می‌توانستیم از فروش محصولاتی که مشتریان درخواست کرده بودند و ما به دلیل وجود نداشتن مواد اولیه قادر به تهیه‌ی آن نبودیم را نیز با علامت منفی در حاصل جمع در نظر بگیریم. هم چنین قیمت مواد اولیه و محصولاتی که تاریخ انقضای آن‌ها در دقیقه گذشته به پایان رسیده ولی مصرف نشده‌اند را با علامت منفی در جمع قرار می‌دهیم.

۷. نحوه‌ی انتقال از یک حالت به حالت دیگر:

در پایان هر دقیقه باید حالت را به روز رسانی کنیم. حالت ما شامل دو ماتریس و یک عدد بود. که آن عدد به سادگی به ازای هر دقیقه یک بار افزایش می‌یابد.

ماتریس مربوط به مواد اولیه در هر روز به این صورت تغییر می‌یابد که در ستون اول که مربوط به مقدار ماده‌ی اولیه موجود در انبار است، به اندازه‌ی مقدار مصرف شده در دقیقه گذشته به روز رسانی می‌شود و ستون دوم نیز هر عدد موجود در لیست به اندازه‌ی یک دقیقه کاهش می‌یابند. و قیمت ماده‌ی اولیه نیز ممکن است با توجه به تغییرات بازار در دقیقه گذشته تغییر کرده یا ثابت بماند.

۸. چرا MDP است ؟ :

یک فرآیند تصمیم گیری مارکوف شامل یک چهارتایی مرتبط است که شامل مجموعه حالات و مجموعه عمل ها و یک تابع احتمال است که در یک حالت خاص با انجام یک عمل مشخص با چه احتمالی به حالت بعدی می‌رویم و هم چنین یک تابع پاداش که در یک حالت خاص با انجام یک عمل مشخص به چه میزان پاداش دریافت می‌کنیم.

خاصیت مارکوف بیان می‌کند که وقتی در یک حالت قرار داریم با فرض انجام یک عمل مشخص، حالت بعدی و پاداش دریافتی کاملا مستقل از حالت ها و اعمال ما در زمان های گذشته باشد. که با توجه به توضیحاتی که در بخش قبل ارائه شد این ویژگی به طور کامل برقرار است.

هم چنین مجموعه حالت ها و اعمال می‌توانند محدود یا نامحدود باشد که با توجه به عدم فرض مسئله می‌توانیم هر کدام را در نظر بگیریم. اگر بخواهیم از مجموعه‌ی محدود استفاده کنیم باید برای قیمت ها سقف و گام تعیین کنیم هم چنین برای خرید اولیه و تولید محصول که این فرض ها محدودیت زا هستند پس از فرض نامحدود بودن مجموعه حالت و عمل استفاده می‌کنیم.

با فرض در اختیار داشتن توزیع افزایش قیمت ها و هم‌چنین داشتن توزیع سفارشات مشتریان و هم چنین با دانستن تاریخ انقضای مواد اولیه و محصولات کارخانه می‌توانیم تابع احتمال مربوط به گذار از یک حالت به حالت دیگر و پاداش دریافتی را نیز محاسبه کرد. یا اینکه می‌توانیم دینامیک محیط را ناشناخته در نظر بگیریم. که در این صورت عامل باید آن را پیدا کند.