گزارش تمرین شماره سه

نام و نام خانوادگی	امیرمحمد رنجبر پازکی
شماره دانشجویی	۸۱۰۱۹۹۳۴۰

سوال 1 – مساله بورس

برای حل این مساله ابتدا نیاز است تا مساله را به صورت یک MDP تعریف کنیم. برای این کار باید اجزای مساله تعریف شود و سپس، به پیاده سازی پرداخته می‌شود.

مهم‌ترین جز در این مساله حالت(state) است. حالت برای این مساله به صورت مجموعه سرمایه فرد و قیمت شرکت‌های B، C و D تعریف می‌شود.

عمل agent سه حالت دارد. یا فرد هیچ سهام نمی‌خرد یا یک سهام یک شرکت را می‌خرد یا دو سهم از دو شرکت را می‌خرد.

پس از هر عمل، با احتمالی به یک حالت دیگر می‌رویم که ممکن است ناشی از تغییر قیمت یک سهام باشد یا بر مبنای تغییر قیمت خریداری شده میزان سرمایه ما نیز تغییر کند. احتمال رفتن به هر state نیز بر مبنای احتمال تغییر قیمت سهام‌ها تعریف می‌شود که تابع احتمال انتقال را می‌سازد. جایزه در این مساله برابر سود یا ضرری است که از معامله سهام داشته‌ایم.

الف) برای پیاده‌سازی این سوال یک environment به نام MDPEnvironment در فایل mdp_env.py نوشته‌شده‌است. این عامل در ابتدا همه حالت‌ها را تولید می‌کند و عملا تابع انتقال را می‌سازد. در حقیقت، next_states بیانگر احتمال حالت بعدی و reward به شرط بودن در حالت فعلی و عملی a است. سپس، عامل مورد نظر با استفاده از تابع get_proability_and_reward محیط با دادن حالت و عمل مورد نظر می‌تواند شناسه حالت بعد، جایزه و احتمال انتقال را بگیرد.

عامل موردنظر در stock_agent.py با نام StockAgent پیاده‌سازی شده‌است. این عامل با استفاده از iterate_policy حلقه‌ی اصلی الگوریتم را اجرا می‌کند. هر حلقه شامل دو بخش evaluate_policy و improve_policy است که به ترتیب به ارزیابی ارزش حالات با استفاده از سیاست فعلی و بهبود سیاست با استفاده از ارزش حالات و اعمال مختلف به‌دست‌می‌آید. لازم به ذکر است برای سیاست نیز یک آرایه یک بعدی در نظر گرفته‌شده‌است چراکه سیاست بهینه به صورت greedy است و عمل بهینه احتمال ۱ و باقی احتمال صفر دارند.

پس از این که اعمال بهینه تغییری نکنند، سیاست پایدار شده‌است و بنابراین، می‌توان سیاست نهایی را دید. سیاست نهایی با استفاده از discount factor برابر ۰.۹ و با حالت اولیه گفته‌شده(دارایی: ۲۰ دلار، ارزش b: ۵ دلار، ارزش c: ۱۵ دلار، ارزش d: ۱۰ دلار)، عمل B است. خروجی کد در زیر دیده‌می‌شود.

```
Iteration #: 11
Delta: 0.0004370701672673505
Action values:
 [7.139897574225592, 10.88248745963236, 7.13970428143842, 5.26713852892011, 10.882180829059234, 9.010938121366536, 5.26698016651593]
----------------------------------------
Optimal Action:
B
```

همانطور که دیده‌می‌شود، بهترین ارزش عمل به ازای خرید سهام **B** به دست‌آمده‌است. خرید **B** و **C** نیز عمل بدی نبوده‌است و بسیار نزدیک به این عمل ارزیابی شده‌است. همانطور که در نتیجه دیده‌می‌شود، این نتیجه طی ۱۱ **iteration** به‌دست‌آمده‌است.

ب) نتایج زیر به ازای discount factor برابر ۰.۲، ۰.۵، ۰.۷ و ۰.۹ به‌دست‌آمده‌است.

- Discount factor: 0.2

```
Iteration #: 14
Delta: 1.7763568394002505e-15
Action values:
 [0.5137459151673545, 4.263745792392325, 0.5137459151494875, -1.3612541616071883, 4.263745792392309, 2.3887457923966857, -1.3612540333176777]
----------------------------------------
Optimal Action:
B
```

همانطور که در تصویر بالا مشاهده می‌شود در ۱۴ iteration نتیجه به‌دست آمده است و عمل B بهینه انتخاب شده‌است. همانطور که می‌بینید، ارزش اعمال به دلیل آینده نگری کمتر کمتر است.

- Discount factor: 0.5

```
Iteration #: 8
Delta: 1.5752487048104058e-06
Action values:
 [1.7117631515804455, 5.461760078947664, 1.7117631453211315, -0.16323786716205174, 5.461760066409002, 3.586760168979847, -0.16323581948064134]
----------------------------------------
Optimal Action:
B
```

همانطور که در تصویر بالا مشاهده می‌شود در ۸ iteration نتیجه به‌دست آمده است و عمل B بهینه انتخاب شده‌است. همانطور که می‌بینید، ارزش اعمال به دلیل آینده نگری بیشتر نسبت به حالت قبل بیشتر است.

- Discount factor: 0.7

```
Iteration #: 10
Delta: 3.205185028054558e-05
Action values:
 [3.2594451139014122, 7.009390104251787, 3.259444050167745, 1.3844535276910224, 7.009387448319545, 5.13441389615141, 1.3844579114229214]
----------------------------------------
Optimal Action:
B
```

همانطور که در تصویر بالا مشاهده می‌شود در ۱۰ iteration نتیجه به‌دست آمده است و عمل B بهینه انتخاب شده‌است. همانطور که می‌بینید، ارزش اعمال باز بیشتر شده‌است. همچنان، انتخاب B و C همزمان نیز اتفاق بدی نیست.

- Discount factor: 0.9

```
Iteration #: 11
Delta: 0.0004370701672673505
Action values:
 [7.139897574225592, 10.88248745963236, 7.13970428143842, 5.26713852892011, 10.882180829059234, 9.010938121366536, 5.266980166515931]
----------------------------------------
Optimal Action:
B
```

همانطور که در تصویر بالا مشاهده می‌شود در ۱۱ iteration نتیجه به‌دست آمده است و عمل B بهینه انتخاب شده‌است. همانطور که می‌بینید، ارزش اعمال باز بیشتر شده‌است. همچنان، انتخاب B و C همزمان نیز اتفاق بدی نیست.

سرمایه گذار هر چه discount factor بزرگ‌تری داشته باشد، بلندمدت تر نگاه می‌کند. در اینجا به دلیل ذات ثابت مسئله(احتمال سود و ضرر ثابت) و برتری نسبی یک سرمایه‌گذاری(B) نسبت به باقی سیاست بهینه تغییر نمی‌کند اما هر چه آینده‌نگری بیشتر باشد(discount factor)، تصمیم دقیق‌تر است و اختلاف تصمیم‌های نزدیک به هم بیشتر می‌شود و در نتیجه، با اطمینان بیشتری می‌توان خرید سهام B را نسبت به B و C توجیه کرد. در تصمیم گیری کوتاه مدت(discount factor کوچک) این مقادیر به یکدیگر نزدیک‌ترند و تصمیم گیری سخت‌تر است. همچنین، ارزش گذاری منفی به ازای بعضی اعمال دیده‌می‌شود که در حالت دید بلند مدت این مسئله حل شده‌است و این نشان می‌دهد که سرمایه گذاری در حالت کلی و در بلند مدت احتمالا سود خود دارد.

از طرفی تا جایی با افزایش discount factor زمان تصمیم‌گیری و انتخاب سیاست بهینه کاهش می‌یابد چراکه اعمال ارزش‌های خود را مطمئن‌تر پیدا می‌کنند و تغییرات کمتری وجود دارد ولی از جایی به بعد تعداد

iterationهای افزایش می‌یابد چراکه عامل بلند مدت‌تر نگاه می‌کند و به همین دلیل، ممکن است آینده‌نگری سیاست را بیشتر تغییر بدهد.

در متغیر policy عامل، عمل بهینه به ازای تک تک حالت وجود دارد. با پیش روی از روی آن می‌توان مختلف رسیدن به خواسته مورد نظر را مورد بررسی قرار داد. (چرا که انتقال به حالات مختلف احتمالاتی است.)

سوال ۲ – کارخانه مواد غذایی

در این سوال حالت را به صورت زیر تعریف می‌کنیم:

- ۵۰۰ عدد برای حجم هر ماده‌ی اولیه: مواد اولیه در طول کار کارخانه به مواد غذایی تبدیل می‌شوند و از آن‌ها کاسته می‌شود یا با خریدن آن‌ها، به مقدارشان افزوده می‌شود.
- ۱۰۰ عدد به عنوان تعداد هر نوع ماده‌ی غذایی: مواد غذایی از روی مواد اولیه ساخته می‌شوند و در انبار ذخیره می‌شوند تا در هنگام درخواست مشتری، آن‌ها برای مشتری ارسال شوند.
- ۵۰۰ عدد به عنوان میزان تقاضا(محبوبیت) هر ماده‌ی اولیه توسط مشتریان: هر چه خدمات رسانی به مشتریان بیشتر شود، میزان تقاضای کلی بالاتر می‌رود. در مقابل، اگر مشتری جنسی را بخواهد که موجود نباشد،نارضایتی حاصل برای او تقاضا را کم می‌کند و حتی ممکن است با پراکنده‌کردن نظر خود در شبکه‌های اجتماعی میزان تقاضای کل را تحت تاثیر قرار دهد. همچنین، میزان تقاضا بر مبنای عوض‌شدن ماه مورد نظر برای هر کالا نیز باید تغییر کند. چرا که درخواست مواد غذایی در ماه‌های مختلف متفاوت بوده و به همین دلیل، تقاضا از ماه جاری اثر می‌گیرد. این میزان محبوبیت هر ماده غذایی روی مواد اولیه آن به نسبت حجم مورد استفاده تقسیم می‌شود و محبوبیت ماده‌ی اولیه را می‌سازد.
- یک متغیر به نام is_first_month که نشان‌دهنده قرار داشتن در اول ماه است.

اعمال، نحوه‌ی انتقال از یک حالت به حالت دیگر و پاداش دریافتی به ازای هر عمل در این محیط به صورت زیر است: (لازم به ذکر است که برخی اعمال به صورت رویداد هستند و در زمان مشخصی اتفاق میافتند.)

1. ۵۰۰ عمل به ازای خرید هر نوع ماده‌ی اولیه: مواد اولیه نیاز به تامین دارند تا بتوان با استفاده از آن‌ها مواد غذایی را تولید کرد. تامین این مواد تنها در اول ماه رخ می‌دهد. یعنی این عمل تنها در حالت اول ماه احتمال رخدادن دارد. همچنین، پس از خرید حالت محیط عوض می‌شود و متغیر is_first_month صفر می‌شود.
   با خرید مواد اولیه حجم آن‌ها در حالت(state) عوض می‌شود. در تامین مواد اولیه به میزان قیمت در حجم مواد اولیه هزینه باید کرد که در این حالت پاداش منفی می‌گیریم. اگر ماده‌ای شرایط سخت برای تامین قرار داشته‌باشد، پاداش منفی حاصل از خرید آن را نصف حساب می‌کنیم چرا که تامین آن نسبت به مواد دیگر هزینه بیشتری دارد و راحت نیست.
   لازم به ذکر است که خرید مواد اولیه به میزانی به حالت اضافه می‌کند و همواره، مقدار مشخصی خرید انجام نمی‌شود. این میزان خرید وابسته به میزان تقاضای هر ماده‌ی اولیه است. احتمال انتقال به حالت‌ها بر مبنای عمل خرید بر مبنای همین میزان تقاضا تعیین می‌شود.
2. ۱۰۰ عمل به ازای تولید هر نوع ماده‌ی غذایی: مواد غذایی با استفاده از مقداری از مواد اولیه تولید می‌شوند. به همین دلیل، برای تولید یک عدد ماده غذایی به میزان مورد نیاز از مواد اولیه مورد استفاده کم می‌شود و در حالت اعمال می‌شود. همچنین، به تعداد مواد غذایی اضافه می‌شود چرا که تولید شده‌اند. در صورت فرض داشتن، انبار می‌توان حجم انبار را در حالت آورد و در صورت نداشتن، فضا پاداش منفی داد که برای سادگی این فرض در نظر گرفته‌نشده‌است. عمل تولید یک عمل deterministic است و با احتمال یک ماده‌ی اولیه به همان میزان کاسته و به تعداد ماده غذایی افزوده می‌شود. (با یک سطح بیشتر پیچیدگی می‌توان میزان تولید را احتمالاتی و برخواسته از توزیعی درنظر گرفت.)

3 . اعمال مربوط به تقاضا و فروش مواد غذایی: این عمل هنگامی رخ می‌دهد که مشتری از یک ماده غذایی تعدادی سفارش می‌دهد. در صورت موجود بودن این تعداد کالا، از تعداد کالا کاسته می‌شود و به میزان قیمت آن‌ها به ما پاداش داده‌می‌شود. در این حالت که سفارش موفقیت آمیز بوده‌است، به احتمالی میزان تقاضا افزایش می‌یابد. فرض کنید این افزایش نیز احتمالاتی و برخواسته از یک توزیع نرمال با میانگین ۱۰ و واریانس ۱ باشد. این افزایش تقاضا می‌تواند به دلیل تبلیغات مثبت یک مشتری و بیان رضایت او در شبکه‌های اجتماعی و غیره باشد. اگر ماده غذایی درخواستی موجود نباشد، نارضایتی برای مشتری ایجاد می‌شود. به دلیل این که طبق prospect theory ما انسان‌ها loss averse هستیم، ضریبی باید در مجازات عدم وجود ماده غذایی ضرب شود. این ضریب را ۲ در نظر می‌گیریم. یعنی، اگر مشتری را از دست بدهیم به میزان قیمت سفارش او نه تنها ضرر کرده‌ایم بلکه به دلیل نارضایتی ایجادشده، دوبرابر این میزان پاداش منفی(مجازات) در نظر می‌گیریم. در این حالت که سفارش موفقیت آمیز نبوده‌است، به احتمالی میزان تقاضا کاهش می‌یابد. فرض کنید این کاهش نیز احتمالاتی و برخواسته از یک توزیع نرمال با میانگین ۲۰ و واریانس ۱ باشد. این نارضایتی می‌تواند به دلیل تبلیغ منفی مشتری در شبکه‌های اجتماعی یا برای دوستانش باشد. در نتیجه، این عمل نیز به صورت احتمالاتی انتقال به سایر حالت را انجام می‌دهد و این احتمال متناسب با وجود یا عدم وجود کالا و میزان افزایش یا کاهش تقاضا و احتمال آن‌هاست.

4 . اعمال مربوط به سرکشی به انبار: ۶۰۰ عمل در این بخش تعریف می‌کنیم. به این صورت که در پایان هر ماه، یک سرکشی به هر ماده اولیه(۵۰۰ تا) و یک سرکشی به هر ماده غذایی(۱۰۰ تا) انجام می‌شود. در هر عمل در صورت گذشتن از میزان ماندگاری مواد آن حجم ماده دور ریخته می‌شود و از حجم ماده در حالت کاسته می‌شود. همچنین، به دلیل این دور ریختن به میزان قیمت ماده دور ریخته شده پاداش منفی(جزا) دریافت می‌شود. این انتقال نیز به صورت deterministic اجرا می‌شود چرا که حالت دیگری وجود ندارد.(اگر احتمال خرابی داشته باشیم، می‌توان با یک احتمالی مواد را نگه داشت یا حجم مشخصی را دور ریخت. برای ساده سازی این فرض در نظر گرفته‌نشده‌است.)

5 . عمل رسیدن به سر ماه: پس از رسیدن به سر ماه متغیر is_first_month به عنوان پرچم در حالت ۱ می‌شود و تغییر حالت با احتمال ۱ رخ می‌دهد. لازم به ذکر است که مابقی المان‌های حالت به جز میزان تقاضای مواد اولیه ثابت می‌مانند.
در هر ماه میزان تقاضای هر ماده غذایی تغییر می‌کند. فرض می‌کنیم میزان تقاضای هر ماده‌ی غذایی با یک ضریب بین ۱ تا ۱۰ تعریف می‌شود که بر مبنای داده قدیمی(عملکرد شرکت در سال‌های قبل) این ضرایب محاسبه شده‌است. پس از رسیدن به سر ماه، به ازای هر ماده‌ی غذایی میزان محبوبیت(تقاضا) مواد اولیه به کار رفته در آن را تغییر می‌دهیم. به این صورت که، میزان تقاضای هر ماده‌اولیه را به میزان تقاضای ماه قبل ماده‌غذایی تقسیم و در میزان تقاضای این ماه ماده‌ی غذایی ضرب می‌کنیم. این کار را به ازای‌همه مواد اولیه‌ی همه مواد غذایی تکرار می‌کنیم و میزان تقاضا(محبوبیت) جدید مواد اولیه در ماه جدید را محاسبه می‌کنیم و در حالت اعمال می‌کنیم تا بتوان میزان خرید مواد اولیه را بر مبنای تقاضا در ماه جدید محاسبه کرد. پس، تغییر حالت با تغییر پرچم و میزان تقاضای مواد اولیه صورت می‌پذیرد. این تغییر نیز به صورت deterministic است. با این کار پاداشی دریافت نمی‌کنیم. (می‌توان فرض کرد که محبوبیت در هر ماه نیز یک توزیع احتمالاتی دارد که در آن صورت این تغییر حالت به صورت احتمالاتی در می‌آید که برای ساده‌سازی در نظر گرفته‌نشده‌است.)

حال که حالات، اعمال، نحوه‌ی انتقال از یک حالت به حالت دیگر به همراه احتمال آن و پاداش دریافتی مشخص شد ما یک مدل MDP برای این مسئله داریم که می‌توانیم با روش policy iteration عمل بهینه را به ازای بودن در هر حالتی محاسبه کنیم.

این مسئله یک مدل MDP است چرا که دنیا از حالاتی تشکیل شده‌است و تعداد این حالت بسیار بیشتر از یک است. در هر حالت ما عملی انجام می‌دهیم که منجر به تغییر حالت و گرفتن پاداشی می‌شود. پاداش متوسط هر عمل در هر حالت و احتمال جابجایی از هر حالت به ازای عمل با یکدیگر متفاوت است و این ذات این مسئله است

که کاملا یک مدل MDP را می‌پذیرد و المان‌های اصلی آن را دارد(مجموعه‌ی حالات(States)، مجموعه‌ی اعمال(Actions)، تابع انتقال از حالتی به حالت دیگر($P_{ss'}^{a}$)، پاداش متوسط انتقال($R_{ss'}^{a}$)). در این مسئله به طور پیوسته در حال زندگی ، تصمیم، پاداش و جزا و حرکت از حالتی به حالت دیگر هستیم.