به نام خدا

تمرین سوم

درس یادگیری تعاملی

پاییز 99

محمد امین حق پناه(mdan.hagh@gmail.com) –روح‌الله ابوالحسنی (r.abolhasani@ut.ac.ir)

سوال 1

هدف از این مساله آشنایی با الگوریتم Policy Iteration و یادگیری تقویتی در فضای چند حالته است.

در یک بازار بورس، سه شرکت سهام خود را عرضه کرده‌اند. یک نفر می‌خواهد با سرمایه‌گذاری صحیح در این بورس، هرچه سریعتر سرمایه اولیه خود را به مقدار هدف ۱۰۰ دلار برساند. سرمایه اولیه این شخص ۲۰ دلار است. در این بورس، قیمت هر سهم از این سه شرکت می‌تواند بصورت احتمالاتی بیشتر شود، کمتر شود یا ثابت بماند. افزایش یا کاهش قیمت هر سهم نسبت به قیمت روز گذشته خود ۵ دلار است. این احتمال‌ها برای این سه شرکت در جدول زیر آمده است.

نام شرکت	احتمال افزایش قیمت	احتمال کاهش قیمت	احتمال ثابت ماندن قیمت
B	0.4	0.3	0.3
C	0.1	0.1	0.8
D	0.2	0.7	0.1

قیمت سهام این سه شرکت در روز اولیه عرضه در جدول زیر مشخص شده است. sid[i] برابر با iامین رقم شماره دانشجویی شما از سمت راست هست. درصورتیکه قیمت اولیه بدست‌آمده برای هر شرکت بر ۵ بخش‌پذیر نبود، آن را به نزدیکترین عدد بخش‌پذیر بر ۵ گرد کنید.

نام شرکت	قیمت اولیه
B	5+sid[1]
C	10+sid[2]
D	5+sid[3]

در این بازار، قیمت هر سهم می‌تواند حداقل ۵ دلار و حداکثر ۵۰ دلار باشد. در صورتیکه قیمت یک سهم به کمترین حد خود برسد، به احتمال 0.25 در همان قیمت باقی می‌ماند و با احتمال 0.75 پنج دلار افزایش می‌یابد. در صورتیکه قیمت یک سهم به بیشترین حد خود برسد، با احتمال 0.25 در همان قیمت باقی می‌ماند و با احتمال 0.75 پنج دلار کاهش می‌یابد.

روش سرمایه‌گذاری در این بازار بدین صورت است که فرد در هر روز می‌تواند سه نوع عمل انجام بدهد:

1. هیچ کاری نکند! یعنی هیچ سهامی نخرد.
2. یک سهم از یکی از شرکت‌ها بخرد.
3. دو سهم از دو شرکت مختلف بخرد.

سرمایه‌گذار باید در ابتدای روز تصمیم خود را عملی کند. پس از انتخاب عمل خود، بازار شروط به تلاطم(!) می‌کند و در انتهای روز، به میزان قیمت جدید سهام یا سهام‌های خریداری شده، پول به حساب سرمایه‌گذار واریز می‌شود. مثلا اگر در ابتدای یک روز سرمایه‌گذار یک سهام به قیمت ۲۰ دلار بخرد و قیمت آن سهام در انتهای روز ۱۵ دلار بشود، سرمایه‌گذار در پایان آن روز ۵ دلار ضرر کرده است.

الف) سیاست بهینه را با روش policy iteration و به ازای discount factor=0.9 بیابید.

ب) الگوریتم بخش الف را به ازای چهار مقدار مختلف discount factor اجرا کرده و نتایج را تحلیل کنید. چه تفاوت‌هایی میان رفتار سرمایه‌گذار به ازای این چهار مقدار وجود دارد؟

سوال ۲

هدف از این مساله طراحی یک مدل MDP است.

یک کارخانه‌ی تولید مواد غذایی قصد دارد تا با استفاده از روش‌های هوش مصنوعی تولیدش را بهینه‌تر کرده و به فروش و سود بیشتری دست پیدا کند. در نهایت هدف این است که یک عامل کامپیوتری مدیریت خرید مواد اولیه و تبدیل آن‌ها به مواد غذایی و فروش را انجام دهد.

این کارخانه قصد دارد تا تعدادی متخصص را گرد هم آورده و مسئله را به صورت حدودی و نه دقیق برای آن‌ها شرح دهد، چرا که تمام جزئیات مسئله برای خود آن‌ها هم هنوز مشخص نیست. شما نیز به عنوان یک متخصص در این گردهمایی حضور دارید و قصد دارید به صاحبان کارخانه نشان دهید که شایستگی بیشتری از سایر متخصصین برای حل این مسئله دارید!

فرض کنید این کارخانه توانایی تولید ۱۰۰ نوع ماده‌ی غذایی مختلف را داشته و برای تولید آن‌ها لازم است تا مجموعا ۵۰۰ نوع ماده‌ی اولیه‌ی مختلف در اختیار داشته باشد. از ویژگی‌های مواد اولیه می‌توان به قیمت، ماندگاری، حجم، سهولت در تهیه (مثلا تحریم باشد یا نه) و ... اشاره کرد. به کمک این مواد اولیه نیز می‌توان مواد غذایی متنوعی تولید کرد و به دست مشتری رساند. هر نوع ماده‌ی غذایی با استفاده از تعدادی ماده‌ی اولیه با حجم‌های متفاوت قابل تولید است و از ویژگی‌های آن می‌توان به قیمت و ماندگاری اشاره کرد. برای ساده‌سازی مسئله، زمان مورد نیاز برای تبدیل ماده‌ی اولیه به ماده‌ی غذایی نهایی را صفر در نظر بگیرید. بنابراین، کارخانه هر زمان که اراده کند، در صورت داشتن ماده‌ی اولیه به مقدار کافی می‌تواند یک ماده‌ی غذایی را در لحظه تولید کند.

کارخانه مواد اولیه را تنها در روز اول ماه می‌تواند بخرد و اگر وسط ماه متوجه شود که به یک ماده‌ی اولیه نیاز جدی دارد، مجبور است تا روز اول ماه بعدی صبر کند. این کارخانه محصولات خود را بدون واسطه و به صورت اینترنتی به مردم می‌فروشد. در هر لحظه مردم می‌توانند وارد سایت کارخانه شده و محصولات موجود را ببینند، سپس سفارش خود را ثبت کنند. برای سادگی، زمان رساندن محصول به مشتریان پس از ثبت سفارش را صفر در نظر بگیرید. همچنین فرض کنید تمامی محصولات کارخانه دارای کیفیت عالی بوده و عدم رضایت مشتری تنها مربوط به موجود بودن یا نبودن مواد غذایی مورد نیاز است.

درخواست‌ها نیز طبیعتا در ماه‌های مختلف متفاوت بوده و بعضی از مواد غذایی تنها در بعضی از ماه‌های سال مشتری دارند. در واقع نیازهای مشتریان در طول زمان متغیر است.

برای حل این مسئله به کمک یادگیری تعاملی، لازم است تا ابتدا یک مدل برای آن ارائه دهیم، سپس با استفاده از این مدل مسئله را حل کنیم. یک مدل MDP برای این مسئله طراحی کنید و حالت‌ها، عمل‌ها، نحوه‌ی انتقال از یک حالت به حالت دیگر و پاداش دریافتی را مشخص کنید و در نهایت توضیح دهید چرا این مدل MDP است.

توجه: این سوال یک پاسخ مشخص ندارد و به صورت مبهم طراحی شده تا به مسائل موجود در دنیای واقعی نزدیک‌تر باشد. هر فرضی که فکر می‌کنید لازم است را ذکر کنید و با استفاده از آن مدل خود را پیشنهاد دهید.

ملاحظات:

- در صورت ابهام یا سوال در مورد تمرین، آنها را در فروم Q&A بنویسید. دستیاران آموزشی در اسرع وقت به سوالات شما پاسخ می دهند. در این صورت بقیه دانشجویان هم از این پرسش و پاسخ استفاده خواهند کرد.
- برای حل سوال اول باید از زبان Python 3.x و پکیج AMALearn استفاده کنید.
- گزارش شما بخش زیادی از نمره این تمرین را دربر می‌گیرد. بنابراین درگزارش خود تمامی مفروضات، جزییات پیاده‌سازی و نتایج خود را توضیح دهید.