به نام خدا

تمرین دوم

درس یادگیری تعاملی

پاییز 99

محمدحسن بیگدلی(mohammadhassanb@gmail.com) – روح‌الله ابوالحسنی(r.abolhasani@ut.ac.ir)

سوال اول

یک 10-armed bandit را مطابق شکل 1-2 کتاب ساتون-بارتو در نظر بگیرید. عملکرد روش Thompson Sampling برای یادگیری در افق 1000 تلاش (trial) را از منظر regret و درصد استفاده از عمل بهینه به ازای 20بار اجرا بررسی کنید. نتایج خود را از منظر درصد استفاده از عمل بهینه با نتیجه روشهای ارایه شده در شکل‌های کتاب به صورت تقریبی و کیفی مقایسه کنید.

سوال دوم

تصور کنید که با اتمام کرونا، کلاس‌های حضوری دانشگاه از سر گرفته شده است و شما در یکی از ترم‌های شلوغ هر روز ساعت ۷:۳۰ صبح کلاس دارید. مسیر شما از انقلاب به سمت امیرآباد است. شما هر روز، با یک برنامه ریزی منظم می‌توانید هر روز ساعت ۶:۵۵ دقیقه در میدان انقلاب باشید.

حال برای آمدن از انقلاب به امیرآباد و دانشکده‌ی فنی دو راه داریم.

1. با تاکسی بیاییم. در این ساعت، تجمع تاکسی‌ها در میدان انقلاب زیاد است و شما به محض اینکه بخواهید با تاکسی بیایید، می‌توانید سوار شده و به سمت دانشکده حرکت کنید.
2. با اتوبوس‌های انقلاب به امیرآباد بیاییم. این اتوبوس‌ها، با یک توزیع احتمال از لحظه‌ی رسیدن شما به میدان، به ایستگاه می‌رسند و شما می‌توانید سوار شوید.

با توجه به ترافیک این ساعت فرق نمی‌کند که شما با تاکسی بیایید و یا با اتوبوس و با هر دو ۲۵ دقیقه در راه خواهید بود. هدف ما این است که هر روز به موقع سر کلاس حاضر باشیم. :)

با توجه به اینکه هزینه‌ی تاکسی بیشتر از اتوبوس است، ابتدا در صف اتوبوس می‌ایستیم. رفتن با اتوبوس برای ما ۲۰۰۰ تومان صرفه جویی دارد. اگر بخواهیم به موقع به کلاس برسیم، اتوبوس باید به موقع بیاید، برای همین اگر در یک روز به نظر می‌رسید که قرار نیست اتوبوس در زمان مورد قبول برای ما بیاید، وقت ارزشمند خود را در صف هدر نمی‌دهیم و با تاکسی به دانشگاه می‌رویم.

الف) یک مدل RL به گونه‌ای پیشنهاد دهید که زمان بهینه صبر کردن در صف را با توجه به توزیع آمدن اتوبوس (که برای شما مجهول است) بدهد. لازم است استدلال کنید که actionها ،reward ها و utility function را در مدل خود با چه منطقی انتخاب کرده اید. این مدل باید به گونه‌ای باشد که بعد از چند روز، شما بتوانید بهترین زمان صبر کردن را باتوجه به utility function انتخاب‌شده بیابید.

برای مدل کردن این مسئله، باید آن را در قالب یک n-armed bandit بیان کنید. درصورتیکه هرگونه فرضی روی شرایط سوال می‌گذارید، آن را در پاسخ خود قید کنید.

ب) مدل را با استفاده از utility function، action ها و reward هایی که در سوال ۱ پیشنهاد دادید، با استفاده از پکیج AMALearn پیاده‌سازی کرده و آن را train کنید تا agent زمان انتظار بهینه در صف را یاد بگیرد.

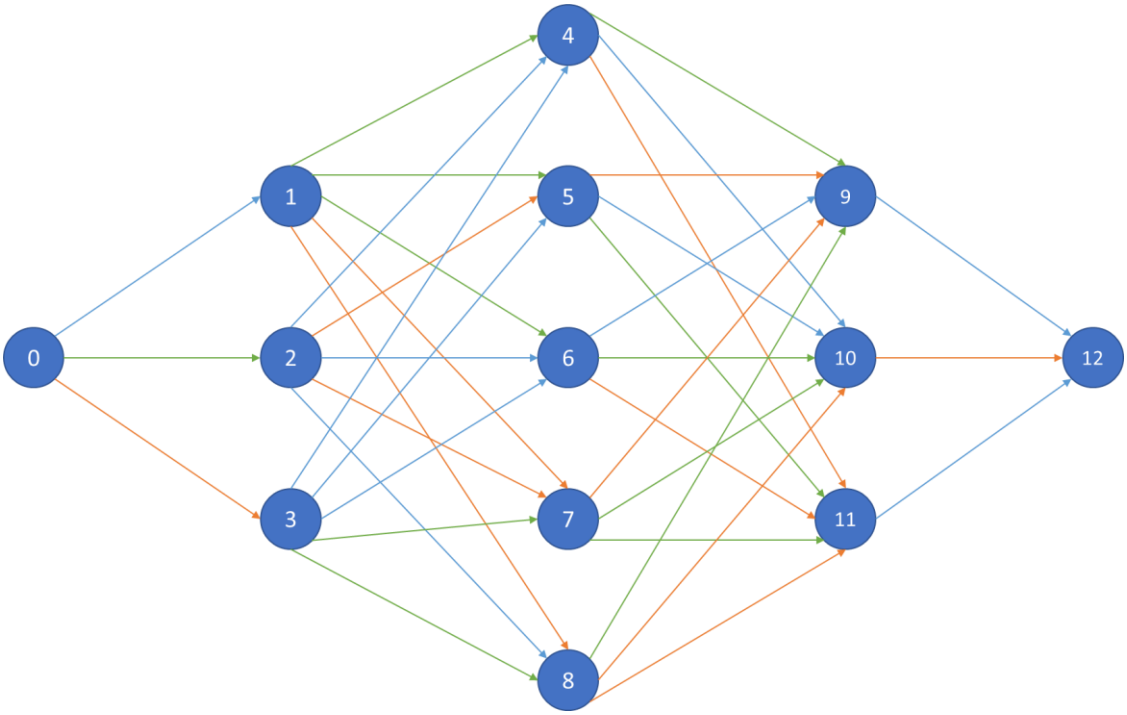
پ) دو agent یکی با سیاست epsilon-greedy و یکی با سیاست UCB توسعه دهید. سپس، رفتار آن‌ها را با هم مقایسه و مزایا و معایب این دو سیاست را بیان کنید. درصورتیکه در بخش قبل، از یکی این سیاست‌ها استفاده کردید، کافیست که عاملی با سیاست دیگر پیاده‌سازی کنید. درصورتیکه از هیچکدام از این دو سیاست در بخش قبل استفاده نکردید، فقط عاملی با سیاست UCB پیاده‌سازی کنید و سپس به مقایسه و تحلیل بپردازید.

ت) (امتیازی) به وسیله‌ی cross validation بهترین learning rate را برای مدلی که در بخش الف توسعه دادید انتخاب کنید.

برای توزیع آمدن اتوبوس در مدل خود می‌توانید از یک تابع گوسی با میانگین۶ و انحراف معیار ۴ استفاده کنید.

سوال سوم

فرض کنید یک شبکه کامپیوتری با ساختاری شبیه به شکل زیر(به جهت فلش‌ها دقت کنید) وجود دارد. قرار هست که بسته‌هایی از گره 0 به گره 12 ارسال شود. هر لینک در این شبکه دارای تاخیر است و ممکن است براساس وضعیت گره‌های میانی، یک لینک موقتا فعال نباشد. درصورتیکه یک گره فعال نباشد، تاخیر دریافت بسته برابر با تاخیر لینک بعلاوه ۳۰ ثانیه زمان ریکاوری گره گیرنده است. فعال بودن یا نبودن هر گره از توزیعی باینری با پارامتر p_i می‌آید که i شماره گره است. تاخیر هر لینک نیز دارای یک توزیع احتمالی مختص آن لینک هست که جزییات آن در ادامه می‌آید.



هدف ما یافتن مسیری است که کمترین تاخیر ارسال را داشته باشد. همچنین می‌خواهیم که بهترین مسیر را با کمترین تعداد آزمایش ممکن بیابیم. لینک‌هایی که در نمودار بالا مشخص شده است، یک‌جهته است. یعنی بسته فقط در جهت فلش قابل ارسال است. همچنین تاخیر رسیدن یک بسته از گره ابتدایی به گره انتهایی یک لینک، بستگی به رنگ آن لینک و شلوغ بودن یا نبودن گره انتهایی دارد. بطور دقیق‌تر، تاخیر هر لینک یک متغیر تصادفی است که از رابطه زیر بدست می‌آید.

$$delay_{i,j} = link_{i,j} + congestion_j$$

در رابطه بالا، $link_{i,j}$ یک متغیر تصادفی است که دربرگیرنده تاخیر لینک براساس رنگ آن است. $congestion_j$ نیز متغیری باینری با پارامتر p_i است. پارامتر p برای هر گره بدان معناست که در صورت رسیدن یک بسته به آن گره، با احتمال p ممکن است که این بسته با تاخیری ۳۰ ثانیه‌ای پردازش شود. توزیع احتمال متغیر تصادفی $link_{color}$ در جدول زیر آمده است. توجه کنید که sid[i] یعنی رقم iام از سمت راست شماره دانشجویی شما. مثلا sid[2] یعنی دومین رقم شماره دانشجویی شما از سمت راست.

رنگ لینک	توزیع احتمالی
آبی	توزیع گوسی با میانگین *sid[1]* و واریانس *sid[2]+0.2*
سبز	توزیع گوسی با میانگین *sid[2]* و واریانس *sid[3]+1*
نارنجی	توزیع گوسی با میانگین *sid[3]* و واریانس *sid[4]+0.5*

همچنین در جدول زیر، پارامتر p_i به ازای گره‌های مختلف تعیین شده است. بدیهی است که این پارامتر برای گره‌های ۰ و ۱۲ معنایی ندارد.

شماره گره	پارامتر p
۱	0.10
۲	0.06
۳	0.15
۴	0.50
۵	0.10
۶	0.15
۷	0.65
۸	0.12
۹	0.20
۱۰	0.05
۱۱	0.45

الف) این مسئله را به فرم n-armed-bandit تبدیل کنید و پاسخ خود را بطور کامل توضیح دهید(مثلا reward و ... چطور تعریف شده‌اند). توجه کنید که جوابی یکتا برای این سوال وجود ندارد و مدلسازی‌های متنوعی برای این سوال می‌توان ارائه کرد.

ب) عاملی بنویسید که با سیاست epsilon greedy بهترین مسیر را با کمترین تعداد آزمایش پیدا کند. رفتار عامل را به ازای epsilon های مختلف تحلیل کنید. نتایج خود را اعم از تاخیر بهترین مسیر و تعداد آزمایش‌های موردنیاز برای یافتن آن، گزارش کنید.

پ) اینبار عاملی توسعه دهید که با روش gradient method بهترین مسیر را در کمترین زمان پیدا کند. عملکرد این عامل را با عاملی که در بخش قبل نوشتید، مقایسه کنید.

ت) بنظر شما برای یافتن پاسخ این مسئله، کدامیک از روش‌هایی که تابحال در درس یادگرفته‌اید مناسب‌تر هست؟ پاسخ خود را توجیه کنید.

ملاحظات

- درصورت وجود ابهام یا سوال در مورد تمرین، آنها را در فروم Q&A بنویسید. دستیاران آموزشی در اسرع وقت به سوالات شما پاسخ می‌دهند. در اینصورت، بقیه دانشجویان هم از این پرسش و پاسخ استفاده خواهند کرد.
- برای حل این سوالات باید از زبان Python 3.x استفاده کنید. همچنین برای ساختن محیط، عامل‌ها و ... باید از پکیج AMALearn استفاده کنید. این پکیج در صفحه درس قابل بارگیری است. برای آشنایی با نحوه کار کردن با آن، ویدئوی ضبط‌شده آن را ببینید.

- سوالات بگونه‌ای طرح شده است که جوابی یکسان ندارد. بنابراین لطفا از خلاقیت و دانش خود برای حل مسائل استفاده کنید.
- برای ارسال این تمرین تا ساعت ۲۳:۵۵ روز ۱۸ آبان فرصت دارید.
- امکان این امر وجود دارد که آزمونک‌ها متکی بر استفاده از کد تولید شده در تکالیف طراحی شوند. لذا پیشنهاد میشود کد تکالیف را به شکل استاندارد توسعه دهید تا استفاده احتمالی از آنها در آزمونکها به سهولت قابل انجام باشد.
- این تکلیف برای تقویت توانایی هر فرد طراحی شده و لذا اجرای آن تک نفره است. بنابراین، همکاری در انجام تکلیف مد نظر نبوده است. از اینکه اخلاق علمی را پاس میدارید متشکریم. برای حفظ حرمت جامعه علمی کوچک خود در این درس، کوشا خواهیم بود.