هدف این تمرین، آزمودن و بکارگیری آموخته‌های شما در مورد مسائل Multi-Armed Bandit است. تمرین از دو بخش سوالات تحلیلی و سوالات پیاده‌سازی تشکیل شده‌است.

در بخش سوالات تحلیلی، سه مساله از مسائل دنیای واقعی انتخاب شده‌اند و برای هرکدام، ارائه یک مدل مبتنی بر مساله Multi-Armed Bandit مورد انتظار است.

در بخش سوال پیاده‌سازی، یک سوال وجود دارد که متشکل از ۵ بخش است. در این سوال، نیاز است تا ابتدا برای یک مساله دنیای واقعی یک مدل مبتنی بر مساله Multi-Armed Bandit ارائه شود. در بخش‌های بعدی نیز به پیاده‌سازی و بررسی الگوریتم‌های مختلف حل مساله Multi-Armed Bandit در شرایط مختلف پرداخته می‌شود.

جزئیات موارد قابل تحویل در صورت هر سوال ذکر شده‌اند.

بخش ۱ – سوالات تحلیلی

برای مسائل زیر، مدلی مبتنی بر مساله Multi-Armed Bandit ارائه کنید. نیاز است تا مجموعه بازوها، پاداش و نحوه پاسخ‌دهی به مساله با توجه به حالات مختلف ورودی به طور دقیق تبیین شود.

۱) فرض کنید در یک باشگاه بدنسازی، سه برنامه تمرینی مختلف به شما ارائه شده‌است که هر برنامه را باید در یکی از روزهای هفته انجام دهید. برنامه مختص هر روز، متشکل از تعدادی حرکت تمرینی است که مربی شما ترتیب اجرای آن‌ها را به عهده خودتان گذاشته‌است (به عنوان مثال، برنامه مختص روز شنبه می‌تواند متشکل از حرکات تمرینی دراز و نشست، شنا و ... باشد). همچنین، به این دلیل که هر حرکت تمرینی می‌تواند وضعیت نرخ سوخت و ساز بدن را تغییر بدهد و همچنین عضلات جدیدی را دخیل کند، ترتیب‌های مختلف اجرای حرکات تمرینی در نهایت منجر به مصرف مقادیر متفاوتی از انرژی می‌شوند. فرض کنید ابزاری در اختیار دارید که می‌تواند میزان انرژی مصرف شده را برای شما اندازه بگیرد.

یک مدل Multi-Armed Bandit برای پیدا کردن بهترین توالی حرکات تمرینی در هر برنامه ارائه کنید.

۲) فرض کنید در مسیری که هر روز برای رسیدن به دانشگاه با خودرو شخصی خود طی می‌کنید چراغ قرمزی وجود دارد که تابلوی زمان‌سنج آن خراب است و مدت زمان باقی مانده برای سبز شدن چراغ را نشان نمی‌دهد. با توجه به بی‌نظمی تقاطع، مدت زمان قرمز ماندن چراغ ممکن است خیلی طول بکشد.

مسیر حرکت شما به گونه‌ای است که در صورت سبز شدن چراغ، پس از گذشت ۱۰ دقیقه به دانشگاه می‌رسید. همچنین در تقاطع مذکور، گردش به راست نیازمند ایستادن پشت چراغ نیست و می‌توانید در صورت تمایل به راست بپیچید و مسیر دیگری را در پیش بگیرید که در این مسیر جدید، برای رسیدن به دانشگاه به ۳۰ دقیقه زمان نیاز است.

یک مدل Multi-Armed Bandit برای تصمیم‌گیری در مورد مدت زمان انتظار پشت چراغ قبل از تغییر مسیر ارائه کنید.

۳) مسیریاب یا Router، وسیله‌ایست که بسته(Packet)های موجود در شبکه اینترنت را به سمت مقاصدشان هدایت می‌کند. مسیریاب‌ها معمولا چند درگاه دارند که هر بسته از یکی از آن‌ها وارد و سپس با توجه به مقصدش از درگاه دیگری گسیل می‌شود. این که بسته‌ها با چه مقاصدی از کدام درگاه مسیریاب گسیل شوند، می‌تواند تاثیر بسیاری در مدت زمان ارسال اطلاعات داشته‌باشد.

فرض کنید یک مسیریاب دارای ۴ درگاه است که امکان دریافت و ارسال بسته‌ها از تمام درگاه‌ها وجود دارد. بسته‌هایی که به این مسیریاب می‌رسند، عموما به مقاصدی در ترکیه، ایران، چین، روسیه و عربستان ارسال شده‌اند. فرض بر این است که به ازای هر بسته‌ای که به مقصدش می‌رسد، یک سیگنال تصدیق(Acknowledgement) از مقصد به مبدا ارسال می‌شود که این سیگنال تصدیق هم از مسیریاب مذکور عبور می‌کند (با اطلاعاتی که از روی بسته‌ها قابل دستیابی است می‌توانیم بفهمیم کدام سیگنال تصدیق مربوط به کدام بسته است).

یک مدل Multi-Armed Bandit برای یادگیری بهترین درگاه برای ارسال بسته‌ها با توجه به مقصدشان ارائه کنید.

بخش ۲ – سوال پیاده‌سازی

یک بانک، برای بخشی از مشتریانش یک طرح تسهیلاتی در نظر گرفته‌است. در این طرح که دانشجویان، کارمندان دولتی و صاحبان مشاغل آزاد را شامل می‌شود، بانک یکی از سه مقدار ۵، ۲۰ و ۱۰۰ میلیون تومان را به عنوان وام به مشتری پرداخت می‌کند و مشتری موظف است در مدت زمان معینی اصل پول به علاوه کارمزد خدمات را به بانک پرداخت کند. مشتری از هرکدام از سه دسته نامبرده که باشد، در صورت دریافت هرکدام از مقادیر موجود برای تسهیلات، با احتمال مشخصی موفق به بازپرداخت مقداری از آن تسهیلات در موعد مقرر می‌شود. این احتمالات در کدهایی که در اختیارتان قرار گرفته، در کلاس Reward داده شده‌است (به عنوان مثال، دانشجویان به احتمال زیاد موفق به بازپرداخت تسهیلات ۱۰۰ میلیون تومانی نمی‌شوند ولی صاحبان مشاغل آزاد اکثرا این تسهیلات را برمی‌گردانند). سیاست‌های بانک باید به گونه‌ای تنظیم گردد که با توجه به توانایی مشتری‌های مختلف در بازپرداخت وام، تسهیلاتی را پیشنهاد دهد که سود بانک بیشینه شود.

در صورتی که مشتری تسهیلات را کامل به بانک برگرداند، بانک می‌تواند از کارمزد دریافتی استفاده ببرد. در غیر این صورت، بانک متحمل ضرری برابر مابه‌التفاوت مبلغ تسهیلات و مقدار بازگردانده‌شده توسط مشتری می‌شود. کارمزد برای تسهیلات ۵، ۲۰ و ۱۰۰ میلیون تومانی به ترتیب ۱۰۰ هزار، ۷۵۰ هزار و ۵ میلیون تومان است.

۱) مدلی بر اساس مساله Multi-Armed Bandit برای بیشینه کردن سود بانک ارائه کنید. مطابق سوالات بخش اول نیاز است تا مجموعه بازوها، پاداش و نحوه پاسخ‌دهی به مساله با توجه به حالات مختلف ورودی به طور دقیق تبیین شود.

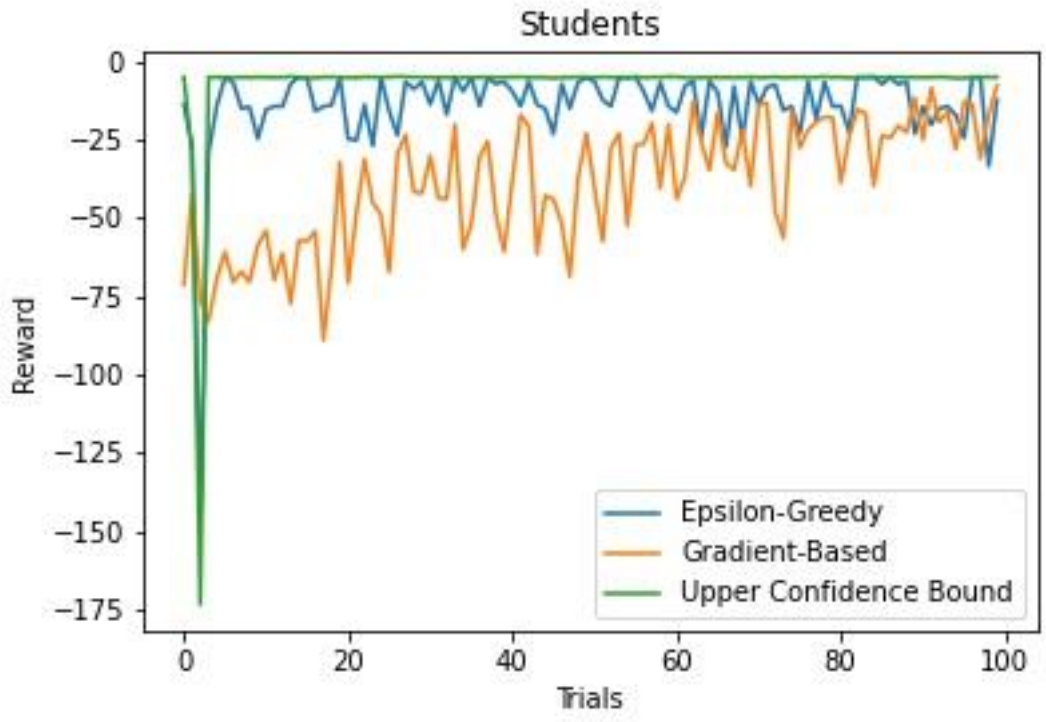
۲) محیط مربوط به مدلی را که در قسمت قبل ارائه کردید، پیاده‌سازی کنید. برای این کار می‌توانید از کدهایی که در اختیارتان قرار داده‌شده استفاده کنید.

توجه داشته باشید که عامل یادگیر به وسیله محیط از مقدار پاداش مطلع می‌شود. همچنین مجموعه بازوها نیز از طریق محیط به عامل داده می‌شود ولی تصمیم‌گیری برای انتخاب از بین آن‌ها برعهده عامل است.

۳) به ازای هرکدام از الگوریتم‌های Epsilon-Greedy، Gradient-Based و Upper Confidence Bound، یک عامل یادگیر Multi-Armed Bandit پیاده‌سازی کنید. برای جلوگیری

از استفاده از کد تکراری در برنامه، می‌توانید از مفاهیم وراثت در شی‌گرایی استفاده کنید. پیاده‌سازی را به شکلی انجام دهید که مقادیر آلفا، بتا و گاما برای محاسبه تابع مطلوبیت (Utility Function، فرض کنید $u = \beta r^{\gamma} + \alpha$) و همچنین هایپرپارامترهای مورد نیاز برای اجرای الگوریتم‌ها (مثل اپسیلون) به عنوان ورودی تابع سازنده به کلاس داده‌شود.

۴) با فرض این که تابع مطلوبیت برابر پاداش دریافتی از بازو باشد، از هرکدام از سه عامل یک نمونه (Instance) بگیرید (مقادیر آلفا، بتا و گاما را طوری تعیین کنید که فرض برقرار شود). اپسیلون را ثابت و برابر ۰.۲، نرخ یادگیری (Learning Rate) را برابر ۰.۰۰۱ و c را برابر ۲ درنظر بگیرید. نمودار میانگین پاداش دریافتی و میانگین مقدار پشیمانی (Regret) در هر ترایال برای ۲۰ بار اجرای هر الگوریتم با ۱۰۰ ترایال را رسم کنید. نمودار نهایی هر گروه مشتریان باید سه خم داشته‌باشد که هرکدام متناظر یک الگوریتم است. به عنوان مثال، نمودار پاداش دریافتی حاصل از اجرای الگوریتم‌های Epsilon-Greedy با اپسیلون برابر ۰.۱، Gradient-Based با نرخ یادگیری برابر ۰.۰۰۰۵ و UCB با c برابر ۴ در محیط متناظر مشتریان دانشجو می‌تواند چیزی شبیه تصویر زیر باشد.



۵) بانک در نظر دارد تا با ارائه تسهیلات مختلف به ۶۰ نفر متقاضی اول، سودده‌ترین تسهیلات برای هرکدام از گروه مشتریان را پیدا کند. با فرض این که در این ۶۰ نفر به تعداد مساوی از هر گروه مشتریان

وجود دارد، نیاز است تا نرخ یادگیری الگوریتم Gradient-Based به گونه‌ای تنظیم شود که در تعداد آزمون مناسب بتواند تسهیلات بهینه را برای هرکدام از گروه‌های مشتریان بیابد. برای این کار، با بررسی ۴ مقدار مختلف نرخ یادگیری، نمودار مربوط به متوسط پاداش دریافتی و متوسط مقدار پشیمانی (Regret) را به تفکیک گروه‌های مشتریان به مانند بخش قبلی (میانگین در ۲۰ اجرا) رسم کنید. انتخاب مقادیر مختلف نرخ یادگیری بر عهده خودتان است، ولی باید در نهایت نرخ یادگیری که به نظرتان بهینه بوده را مشخص کنید (طبیعتا نمودار متناظر این مقدار باید یکی از ۴ خم موجود در نمودارهای نهایی باشد).

نکات پیاده سازی و تحویل

- مهلت ارسال این تمرین تا پایان روز یکشنبه ۸ آبان ماه خواهد بود.
- انجام این تمرین به صورت یک نفره می باشد.
- حجم گزارش معیار تعیین نمره شما نیست، ولی نیاز است تا توضیحات موجود در آن شفاف و کافی باشند.
- از نمودارهای واضح در گزارش خود استفاده کنید. نمودارهایتان حتما روی هر محور و هر خم دارای برچسب واضح باشد.
- کدهایی که برای هر بخش تحویل داده می‌شوند باید قابل اجرا باشند.
- لطفا در گزارش و کدهای خود از تمرین دیگران استفاده نکنید. مشورت و همفکری در مورد سوالات ایرادی ندارد اما اگر شباهت بیش از اندازه در تمرینات مشاهده شود منجر به از دست رفتن نمره تمرین برای تمام افراد خواهدشد.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمائید.

HW2_[Lastname]_[StudentNumber].zip

- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با دستیار آموزشی در ارتباط باشید:

azimpour102@ut.ac.ir
amirali.ataei@gmail.com