به نام او، برای او

تمرین دوم درس یادگیری ماشین

پاییز ۱۳۹۷

سوال ۱.

مسأله‌ی 3-armed bandit را با بازو های زیر درنظر بگیرید:

با فرض $(p,q,l) = (0.5, 0.6, 0.6)$ مساله را با دو سیاست زیر پیاده‌سازی کنید و بازوی بهینه را بیابید. در بازوی اول p احتمال گرفتن پاداش از توزیع $N(60\ ,\ 8)$ در بازوی اول است و $1-p$ احتمال گرفتن پاداش از توزیع $N(-40\ ,\ 8)$ است. متغیرهای q و l نیز همین معنا را در مورد توزیع پاداش بازوهای دیگر دارند.

الف) $\varepsilon - greedy$

ب) $UCB\ 1$

$$arm1: \begin{cases} p: N(60\ ,\ 8) \\ 1-p: N(-40\ ,\ 8) \end{cases}$$

$$arm2: \begin{cases} q: U(40\ ,\ 60) \\ 1-q: U(-40\ ,\ -70) \end{cases}$$

$$arm3: \begin{cases} l: N(20\ ,\ 8) \\ 1-l: U(-10\ ,\ 10) \end{cases}$$

پ) فرض کنید در این مسأله، هرگاه پاداش بازوی انتخاب‌شده مثبت بود، به شما 1000 تومان جایزه می‌دهند و در غیر اینصورت، جایزه‌ای دریافت نمی‌کنید. با توجه به تابع ارزش جدید، الگوریتم یادگیری توسعه دهید که بتواند بازویی که با **احتمال** بیشتری پاداش میدهد را به دست آورد.[1]

ت) با فرض آنکه (p,q,l) بعد از هر k تا $trial$ تغییر کند، چگونه می‌توان این تغییرات را در محیط شناسایی کرد و بازوی بهینه را در تمام مواقع به دست آورد؟[2] (با فرض آنکه k از توزیع $U(30\ ,\ 50)$ می‌آید، (p,q,l) هر کدام از توزیع $U(0,1)$ می‌آیند و عامل یادگیر[3] تمام این اطلاعات را دارد.)

---

[1] برای توضیح بیشتر می‌توانید به مقاله Problem Analysis of Thompson Sampling for the multi-armed bandit از Shipra Agrawal مراجعه کنید.

[2] برای توضیحات بیشتر می‌توانید به مقاله Uncertainty and learning از Peter Dayan & Angelina Yu مراجعه کنید.

[3] Learner Agent

سوال ۲.

به سوالات زیر به صورت تفسیری پاسخ دهید و در صورت امکان، شبه کد بنویسید.

الف) در یک مسأله $n-arm\ bandit$ از ابتدای یادگیری، واریانس پاداش هر یک از بازوها را می‌دانیم. چگونه می‌توانیم از این اطلاعات برای حل مسأله استفاده کنیم؟

ب) در ابتدای یادگیری مسأله $3-arm\ bandit$ ، به ما گفته شده‌است که نسبت میانگین بازوی اول و دوم، k است. یعنی یکی از دو حالت $\mu_{arm_1} = k^*\mu_{arm_2}$ یا $\mu_{arm_2} = k^*\mu_{arm_1}$ برقرار است. اما ما نمی‌دانیم کدام حالت برقرار است. 1- اگر k مشخص باشد، چگونه می‌توان از این اطلاعات برای حل مسأله استفاده کرد؟ 2- اگر k نامشخص باشد چطور؟

سوال ۳.

یک شرکت بازی‌سازی، سه نوع بازی طراحی کرده است و در حال حاضر می‌خواهد تنها یکی از آن‌ها را منتشر کند. افراد بازی نوع اول را با احتمال 0.7 بازی دوم را با احتمال 0.5 و بازی سوم را با احتمال 0.3 می‌برند. طبق تحقیقات انجام‌شده افرادی که قرار است مخاطبان بازی باشند همه با سیاست win stay lose shift بازی می‌کنند و به سه دسته تقسیم می‌شوند: (stay به معنای ادامه دادن بازی و shift خارج شدن از بازی است)

- گروه اول: این گروه اگر ببرند با احتمال 0.5 به بازی کردن ادامه می‌دهند و اگر ببازند با احتمال 0.5 از بازی خارج می‌شوند و دوباره به بازی برنمی‌گردند. این گروه 20% جامعه را تشکیل می‌دهند. این افراد برای سرگرمی و گذراندن وقت بازی می‌کنند!
- گروه دوم: این گروه اگر ببرند با احتمال 0.9 به بازی کردن ادامه می‌دهند و اگر ببازند با احتمال 0.3 از بازی خارج می‌شوند و دوباره به بازی برنمی‌گردند. این گروه 20% جامعه را تشکیل می‌دهند. این افراد میل شدیدی به بازی کردن دارند. ☺️
- گروه سوم: این گروه اگر ببرند با احتمال 0.9 به بازی کردن ادامه می‌دهند و اگر ببازند با احتمال 0.9 از بازی خارج می‌شوند و دوباره به بازی برنمی‌گردند. این گروه 60% جامعه را تشکیل می‌دهند. این افراد تا زمانی که ببرند میل به ادامه‌ی بازی دارند و در صورت باخت علاقه‌ای به ادامه دادن ندارند.

این شرکت از راه نشان دادن تبلیغ به بازیکنان درآمد کسب می‌کند. اگر فردی ببازد و بخواهد به بازی کردن ادامه بدهد باید تبلیغ را مشاهده کند. این شرکت به ازای هر دفعه‌ای که بازی انجام می‌شود و نمی‌تواند تبلیغ نشان دهد هزار تومان ضرر می‌کند و هر دفعه‌ای که تبلیغ نشان می‌دهد 10 هزار تومان سود می‌کند.

الف) با استفاده از شبیه‌سازی نشان دهید اگر هر بازی را 10000 نفر نصب کنند (هر نفر فقط یک بازی را نصب می‌کند) شرکت با منتشر کردن کدام بازی سود بیشتری می‌کند؟ فرض کنید هیچ بازیکنی بیش از 10 مرحله از یک بازی انجام نمی‌دهد. برای هر بازی، بازه‌ی اطمینان 90% سود شرکت را محاسبه و اعلام کنید.

ب) برای نسخه‌ی بعدی بازی شرکت قصد دارد تغییراتی را در بازی ایجاد کند. در صورتی که بازیکنان مرحله‌ای را با موفقیت پشت سر بگذارند، امتیازی را دریافت می‌کنند و در صورت عدم موفقیت هیچ امتیازی نمی‌گیرند. این امتیاز تصادفی است و از یک توزیع نرمال با میانگین 10 و انحراف معیار 3 به دست می‌آید. روانشناسان بر این باور هستند که افراد در این مدل بازی، تعریف برد و باخت را در ذهن خود تغییر می‌دهند. از دید بازیکنان دریافت امتیازی بیشتر از میانگین امتیازی که خودشان تا این مرحله کسب کرده‌اند، برد تلقی شده و در غیر این صورت احساس باخت می‌کنند. با فرض ثابت ماندن ترکیب جامعه، برای سود شرکت چه اتفاقی رخ می‌دهد؟

سوال ۴.

این روزها یکی از راه‌های معرفی و فروش یک کالا تبلیغ از طریق کانال‌های تلگرامی است. فرض کنید دو کانال موجود هستند. برای این‌که تبلیغ خود را در یک کانال قرار دهیم لازم است تا هزینه‌ای را برای آن به مدیر کانال بپردازیم. می‌دانیم که کانال‌های موجود سفارش‌های تبلیغاتی مختلفی دارند، بنابراین هزینه‌ای که در هر بار درخواست و در هر بازه‌ی زمانی معین برای تبلیغ می‌پردازیم با هم متفاوت خواهد بود. به طور کلی برای هر یک از دو کانال سه بازه‌ی زمانی ۲۴-۱۶، ۱۶-۸، ۸-۰ در نظر می‌گیریم. در هر روز به دلیل محدودیت‌ها مجبوریم فقط در یکی از دو کانال و در یکی از سه بازه‌ی زمانی ذکرشده تبلیغ خود را بفرستیم. هدف ما پیدا کردن کانالی‌ست که اگر در یکی از سه بازه‌ی زمانی ذکر شده در آن تبلیغ داشته‌باشیم سود ما بیشینه شود.

سود ما در هر روز از اختلاف بین فروش و هزینه‌ی تبلیغ به دست می‌آید. فروش در یک کانال به دو عامل تعداد بازدیدکننده و تعداد کلیک های روی تبلیغ وابسته است که از طریق رابطه‌ی زیر تعداد کالاهای فروش رفته محاسبه می‌شود.

Purchase= a*number of views + b*number of clicks

در این مسئله ضرایب را به این صورت قرار دهید: a= 0.4 و b= 0.6

برای به دست آوردن سود حاصل از فروش از رابطه‌ی زیر استفاده می‌کنیم:

Benefit= Purchase*P - cost

P نمایانگر هزینه‌ای است که هر نفر به طور متوسط برای خرید از لینک تبلیغ ما پرداخت می‌کند. $Purchase$ میزان فروش ما و $cost$ هزینه‌ای است که برای تبلیغ در آن بازه زمانی باید به مدیر کانال بپردازیم.

در فایل Q4.csv که کنار این سوال آپلود شده است، برای دو کانال ۱ و ۲ که هر کدام به ترتیب 50K و 75K عضو دارند، در هر روز تعداد بازدیدکننده‌ها، تعداد کلیک‌ها و هزینه‌ی قراردادن تبلیغ به ازای هر یک از سه بازه‌ی زمانی مختلف قرارداده‌شده است.

توجه‌کنید که این مسئله یک مسئله‌ی یادگیری تقویتی است و هدف ما این است که با تعداد تجربه‌های کمتر پاسخ را پیدا کنیم.

با توجه به توضیحات، به پرسش‌های زیر پاسخ دهید:

الف) با استفاده از روش $Reinforcement\ Comparison$ به ازای پارامتر $p = 2$ و $p = 8$ (که p همان پارامتر مطرح شده‌ی صورت سوال در محاسبه سود است) به حل مساله بپردازید. در این روش مقدار α=0.1 و β=0.9 در نظر بگیرید. چه تفاوتی بین پاسخ‌ها مشاهده می‌کنید؟ آیا در هر دو حالت الگوریتم به پاسخ بهینه می‌رسد؟ این تفاوت ناشی از چیست؟

ب) نمودار پشیمانی بر حسب تعداد تجربه را رسم کنید و آن را توجیه کنید. به نظر شما مقدار اولیه‌ی α و β باید چه رابطه‌ای با هم داشته باشند؟ (بزرگتر-کوچکتر-مساوی یا مستقل از یکدیگر)

پ) تقریبا پس از چه روزی، مسئله به پاسخ بهینه هم‌گرا می‌شود؟

ت) آیا راهی وجود دارد که یک سیاست حریصانه به پاسخ مسئله برسد؟

لطفا به نکات زیر توجه کنید:

- ✓ حجم گزارش شما به هیچ‌وجه معیار نمره‌دهی نیست، پس لطفا در حد نیاز توضیح دهید.
- ✓ تایپ کردن تمرین‌ها اجباری نیست ولی در صورتی که روی کاغذ می‌نویسید علاوه بر آپلود اسکن در صفحه‌ی درس، برگه‌ی خود را در اولین کلاس درس پس از ددلاین به استاد تحویل دهید.
- ✓ سعی کنید از پاسخ‌های روشن در گزارش خود استفاده کنید و اگر پیش‌فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر کنید.
- ✓ از نمودارهای واضح در گزارش خود استفاده کنید، نمودارهایی که دارای لیبل‌گذاری روشن روی هر محور و همینطور توضیح مناسب باشد.
- ✓ کدهایی که به همراه گزارش تحویل می‌دهید باید قابل اجرا باشد. همچنین توجه کنید که به تمرین بدون گزارش نمره‌ای تعلق نمی‌گیرد.
- ✓ لطفا در گزارش و کدهای خود از تمرین دیگران استفاده نکنید، مشورت و همفکری در مورد سوال‌ها اشکالی ندارد اما اگر شباهت بیش از اندازه در تمرین‌ها دیده شود منجر به صفر شدن نمره خواهد شد.
- ✓ تمام فایل‌ها را در قالب یک فایل zip یا rar در سایت درس بارگذاری کنید.
- ✓ برای پیاده‌سازی تمرین فقط از زبان‌های MATLAB و یا Python می‌توانید استفاده کنید.

موفق و سلامت باشید. :)