به نام خدا

تمرین دوم درس یادگیری تعاملی

بخش اول

پاییز ۱۴۰۰

سوال اول.

در یک مرکز تفریحی یک ماشین شانس ۴ دکمه ای، قرار داده اند.

دکمه اول از توزیع N(a, 1)، دکمه دوم از توزیع N(b, 2)، دکمه سوم از توزیع N(c, 1) و دکمه چهارم با احتمال ۰,۷ از توزیع N(d,2) و با احتمال ۰,۳ از توزیع U(-d, 1) پاداش به بازیکنان میدهد. (دقت کنید که a رقم یکان شماره دانشجویی شما باشد، b=a-2، c=b-1 و d رقم دهگان شماره دانشجویی شما است.)

قسمت اول)

لطفا موارد خواسته شده زیر را به صورت میانگین ۱۰۰۰ ترایال اجرای کل بازی برای ۲۰ بار اجرا، پیاده سازی نمایید و نمودار میانگین پاداش کل را رسم نمایید.

الف) الگوریتم epsilon-greedy با اپسیلون ثابت (۰,۲) الگوریتم یادگیری بر پایه گرادیان و UCB1 را پیاده سازی کنید.

ب) فرض کنید که Mr. Nobody پاداش خود را بر حسب utility function با آلفا برابر ۱، بتا برابر ۱ و گاما برابر ۱ دریافت میکند. حال برای دو مقدار مختلف از هر پارامتر utility بررسی کنید که تاثیر دریافت پاداش با utility های مختلف بر نتیجه ی بخش الف از منظر سرعت یادگیری و سیاست همگرایی چگونه خواهد بود؟

قسمت دوم)

لطفا به موارد زیر به صورت تحلیلی پاسخ خود را ارائه دهید.

الف) تاثیر اپسیلون بر سرعت یادگیری را تحلیل کنید و نتیجه روش اپسیلون ثابت با متغیر را مقایسه کنید.

ب) فرض کنید علاوه بر Mr. Nobody در آن لحظه، ۳ بازیکن دیگر نیز در محیط حضور دارند. در هر دور بازی Mr. Nobody آخرین فردی است که بازی میکند. هر کدام از این ۴ نفر به ترتیب یک بار اقدام به انتخاب یک دکمه برای بازی میکنند. از آنجا که Mr. Nobody فرد زرنگی است، نحوه انتخاب دکمه بازیکنان دیگران را نگاه میکند ولی متاسفانه نمی تواند پاداش دریافتی و عکس العمل دیگر بازیکنان را ببیند. حال به سوالات زیر پاسخ دهید.

۱) **اگر آن سه بازیکن به ترتیب سیاست رفتاری epsilon-greedy با epsilon = ۰,۳ و UCB با c=۲ و راندوم داشته باشند، با توجه به اینکه Mr. Nobody اطلاعاتی نسبت به سیاست دیگران ندارد، شبه کدی ارائه دهید که او چگونه میتواند از رفتار مشاهده شده از آنها برای دریافت پاداش بیشتر و برنده شدن استفاده کند؟ تحلیل کنید که الگوریتم شما در طول زمان یادگیری از کدام بازیکن استفاده ی بهتری می تواند ببرد.**[1]

[1] **برای مطالعه بیشتر می توانید به مقاله** Social Bandit Learning: Strangers Can Help **مراجعه نمایید.**

۲) به نظر شما بهتر است Mr. Nobody از همان ابتدا تقلب کند و یا اجازه دهد مدت زمانی بگذرد؟ پاسخ خود را تحلیل کنید!

۳) اگر به خاطر ضعیف بودن چشم Mr. Nobody ، با خطا بتواند درست تشخیص دهد، آیا تقلب باز هم برای او سودمند خواهد بود؟ صرفا شهود خود را نسبت به تاثیر خطای دید در عملکرد الگوریتم خود بیان کنید.

۴) (بخش امتیازی) فرض کنید سیاست Mr. Nobody برای تصمیم گیری در این مسئله به شکل epsilon-greedy با epsilon برابر با ۰,۲ باشد. الگوریتمی که در بخش ج ارائه دادی پیاده کنید و میانگین پاداش دریافتی از یادگیری فردی و جمعی را مقایسه کنید. (دقت کنید که در این حالت الگوریتم فردی epsilon-greedy با epsilon برابر ۰,۲ باشد) با محاسبه تی-تست ۹۵ درصد بررسی کنید که آیا این دو روش تفاوت معناداری دارند یا خیر.

سوال دوم.

به سوالات زیر به صورت تحلیلی پاسخ دهید و درصورت امکان برای آن شبه کد بنویسید.

الف) یک مسئله multi agent multi armed bandit **را در نظر بگیرید. شما میتوانید رفتار سایر عامل ها را ببینید ولی درکی از پاداش دریافتی آنها ندارید. اما در این مسئله آنها به شما واریانس پاداش هر کدام از بازوها را میگویند. با این اطلاعات چگونه مسئله را حل میکنید؟**

ب) در یک مسئله n-armed bandit**، با گذر زمان، واریانس پاداش های دریافتی به ازای انتخاب بازوها به صورت غیرقابل پیش بینی تغییر می کند. در صورتیکه نرخ یادگیری ارزش اعمال برابر با** $1/m$ **باشد (**m **نشان دهنده تعداد انتخاب بازو است)، چه تاثیری در روند یادگیری عامل خواهد داشت؟**

سوال سوم.

فدراسیون فوتبال ایران در نظر دارد از بین دوشرکت آدیداس و نایکی برای جام جهانی آینده به عنوان تولید کننده لباس تیم ملی ایران انتخاب کند. فدراسیون ۳ طرح برای لباس تیم ملی ایران با رنگ اصلی سفید از شرکت آدیداس و ۳ طرح دیگر با رنگ اصلی قرمز از شرکت نایکی در نظر گرفته است. پیش از نهایی شدن طرح اصلی برای لباس تیم ملی ایران، آنها می خواهند این طرح ها را مورد ارزیابی قرار دهند و به همین دلیل از تعدادی از علاقه مندان به فوتبال میخواهند تا بین ۰ تا ۱۰۰ نمره ای را به هر کدام از این لباس ها در نظر بگیرند. البته برای این کار، هزینه ای را نیز به باید پرداخت کند که با توجه به جنس پارچه، هزینه تیم طراحی و راضی کردن فرد داوطلب، مقدار مشخص خود را دارد.

فدراسیون به دنبال آن است که معیاری برای ارزشمندی لباس اصلی تیم را با توجه به امتیاز هر کدام از افراد و هزینه متحمل شده از برای خرید لباس، برای هر کدام از لباس ها را مشخص کند. معیار ارزش را به صورت زیر تعریف می کنیم:

$$value = 2 * score - cost$$

در فایل Q3.csv برای هر کدام از این طرح ها، امتیاز کاربران و هزینه نهایی آورده شده است. از آنجایی که مسئله ما یادگیری تقویتی است، دنبال آن هستیم تا با حداقل میزان تجربه طرح مناسب را برای لباس تیم ملی ایران پیدا کنیم.

الف) به کمک روش reinforcement comparison برای دو مقدار آلفا برابر با ۰,۱ و بتا برابر با ۰,۹ مسئله را حل کنید .

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

ج) آیا راهی وجود دارد که یک سیاست حریصانه به جواب همگرا شود؟

لطفا به نکات زیر توجه فرمایید:

- ✓ **حجم گزارش شما به هیچ وجه معیار نمره دهی نیست، پس لطفا در حد نیاز توضیح دهید.**
- ✓ **سعی کنید از پاسخ های روشن در گزارش خود استفاده کنید و اگر پیش فرضی در حل سوال در ذهن خود دارید، حتما در گزارش خود آن را ذکر نمایید.**
- ✓ **از نمودارهای واضح در گزارش خود استفاده کنید، نمودارهایتان حتما دارای لیبل واضح روی هر محور و توضیح مناسب باشد.**
- ✓ **کدهایی که به همراه گزارش تحویل می دهید حتما باید قابل اجرا باشند. توجه نمایید که به تمرین بدون گزارش نمره ای تعلق نمیگیرد.**
- ✓ **لطفا در گزارش و کدهای خود از تمرین دیگران استفاده نکنید. مشورت و همفکری در مورد سوال ها اشکالی ندارد اما اگر شباهت بیش از اندازه در تمرین ها دیده شود منجر به صفر شدن نمره خواهد شد.**
- ✓ **تمام فایل ها را در قالب یک فایل zip یا rar در سایت درس بارگذاری کنید.**
- ✓ **حتما فرمت گزارش که در سایت درس قرار داده شده است را رعایت نمایید.**
- ✓ **در صورت وجود هر نوع سوال در رابطه با این تمرین میتونید از طریق آدرس ایمیل alinaghdi8@gmail.com در ارتباط باشید.**

همیشه سلامت باشید ☺