دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

گزارش تمرین شماره 2

درس یادگیری تعاملی

پاییز 1401

امیر حسین بیرژندی

...

810198367

...

فهرست

چکیده

در این تمرین به بررسی مسائل Multi-Armed Bandit می پردازیم. در ابتدا در مسائل تحلیلی مدل مربوط به این موضوع در دنیای واقعی ارائه می‌شود. سپس در مسئله پیاده‌سازی یک مسئله در دنیای واقعی را پیاده‌سازی می‌کنیم.

سوال 1 - سوال تئوری

1)

بازو ها یا همان عمل هایی که می‌توانیم انجام دهیم حرکت های تمرینی موجود در برنامه یک روز است. یعنی مجموعه آن متشکل از حرکت های آن روز است برای مثال مجموعه یک برنامه می تواند به صورت زیر باشد.

Actions = { exercise 1 , exercise 2, … , exercise k}

با توجه به صورت سوال هدف بیشینه کردن مقدار سوخت و ساز بدن ما است در نتیجه پاداش این مسئله میزان انرژی مصرف شده است که توسط آن ابزار خاص اندازه‌گیری می‌شود.

مسئله ما سه برنامه تمرینی مختلف دارد در نتیجه برای یادگیری اینکه در هر برنامه چه توالی‌ای از حرکات بیشترین سوخت و ساز را دارد نیاز به سه agent داریم و می توان بگوییم سه context داریم که هر کدام از 3 برنامه هستند. هر کدام از agent های ما نیاز دارند بدانند که تمرین های مربوط به خودشان با چه توالی بیشترین انرژی را مصرف خواهند کرد. نحوه پاسخ‌دهی ما به مسئله نیز این چنین خواهد بود که با تست کردن ترتیب های مختلف برای هر کشور و دریافت پاداش ها با توجه به سیاستمان اکشن بهینه را پیدا می‌کنیم.

2)

در این مسئله دو بازو یا دو اکشن داریم که متشکل از 1) ایستادن در چهارراه و منتظر چراغ سبز شدن است و 2) تغییر مسیر است.

Actions = {تغییر مسیر دادن , ایستادن و منتظر چراغ سبز بودن }

البته که باید دقت کنیم که این انتخاب بین این دو اکشن در هر لحظه ممکن است صورت گیرد و هر لحظه باید انتخاب کنیم بمانیم یا برویم.

با توجه به صورت مسئله پاداش ما زمان رسیدن به دانشگاه می‌باشد.

در این مسئله به یک agent یادگیر بیشتر نیاز نداریم که این agent در context که همان تقاطع مذکور است قرار دارد. در واقع این agent باید با توجه به تجربه‌هایی که تا کنون داشته یاد بگیرد که تا چه زمانی ارزش دارد پشت چراغ بایستیم و منتظر سبز شدن آن شویم و اگر آن زمان گذشت از مسیر دیگر برویم.

3)

بازو ها یا همان عمل هایی که می‌توانیم در این مسئله انجام دهیم همان 4 درگاه مسیریاب ما هستند یعنی 4 انتخاب داریم و باید بین این 4 تا اکشن با توجه به استیت فعلی اکشن بهینه را انتخاب کنیم.

همچنین پاداش ما در این مسئله مقدار زمانی است که طول می‌کشد سیگنال تصدیق از مبدا به مقصد ارسال شود. در واقع هر چه این زمان کمتر باشد آن درگاه مذکور بهتر عمل کرده است و طبیعتا ریوارد بزرگتری باید دریافت کند.

در این مسئله 5 context یا به تعبیری استیت داریم که همان کشور های مذکور در صورت سوال می‌باشند یعنی ترکیه، ایران، چین، روسیه و عربستان. یعنی ما باید درگاه بهینه در هر کدام از کشور ها را پیدا کنیم و قرار نیست که لزوما درگاهی که برای ایران خوب کار می‌کند برای ترکیه نیز خوب کار کند. نحوه پاسخ‌دهی ما به مسئله نیز این چنین خواهد بود که با تست کردن درگاه های مختلف برای هر کشور و دریافت پاداش ها با توجه به سیاستمان اکشن بهینه را پیدا می‌کنیم.

سوال 2 - سوال پیاده‌سازی

بخش 1

بازوها یا عمل هایی که می‌توانیم انجام دهیم انتخاب بین نوع وام هایی است که می‌توانیم بدهیم که 3 اکشن است اکشن 5 میلیونی، 20 میلیونی و 100 میلیونی است.

پاداش ما در این مسئله مابه‌التفاوت مبلغ تسهیلات و مقدار بازگردانده شده توسط مشتری است.

در این مسئله 3 تا context داریم که نوع مشتری های ما محسوب می‌شوند یعنی یا مشتری یا کارمند یا افراد دارای مشاغل آزاد هستند.

بخش 2 و 3

عوامل و محیط در فایل جوپیتر قرار گرفته‌اند.

برای هر یک از روش های تعامل epsilon greedy ، UCB و Gradient-Bandit یک عامل طراحی شده است. همچنین یک کلاس Action طراحی شده که از طریق محیط به عامل داده می‌شود.

بخش 4

برای اینکه تابع مطلوبیت برابر پاداش دریافتی از بازو باشد مقادیر را به صورت زیر انتخاب می‌کنیم.

$$\alpha = 0 , \beta = 1 , \gamma = 0$$

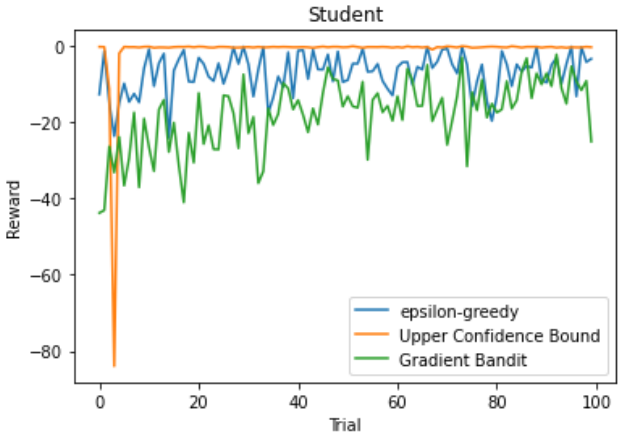
برای محاسبه مقدار متوسط پشیمانی برای هر یک از نوع مشتریان expected value هر یک از 3 اکشن را حساب کرده و مقدار بیشتر را به عنوان حد مطلوب در نظر گرفتیم.

Students -> action 1 -> -0.14

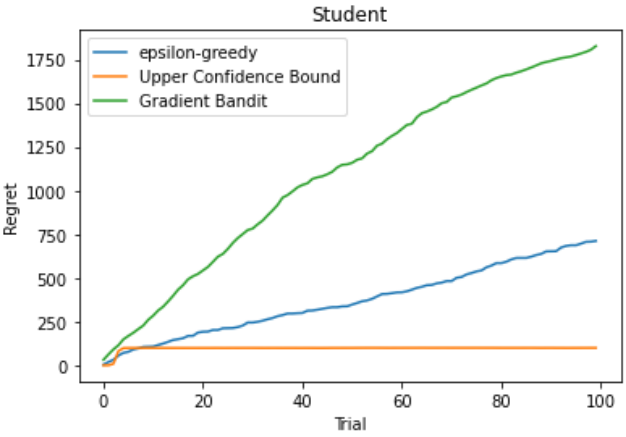
Government Staff -> action 1 -> -0.015

Self-Employed -> action 3 -> -0.6

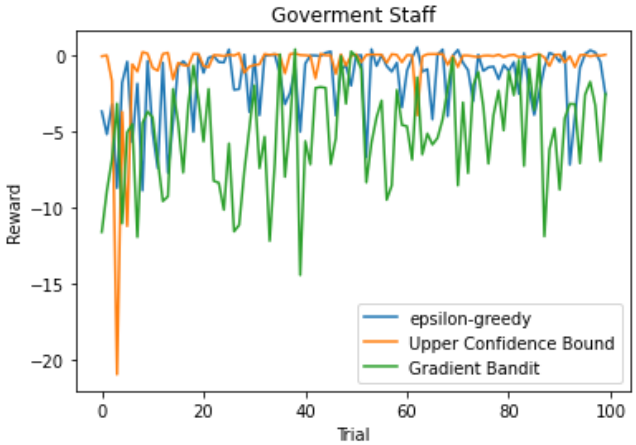
نتایج:



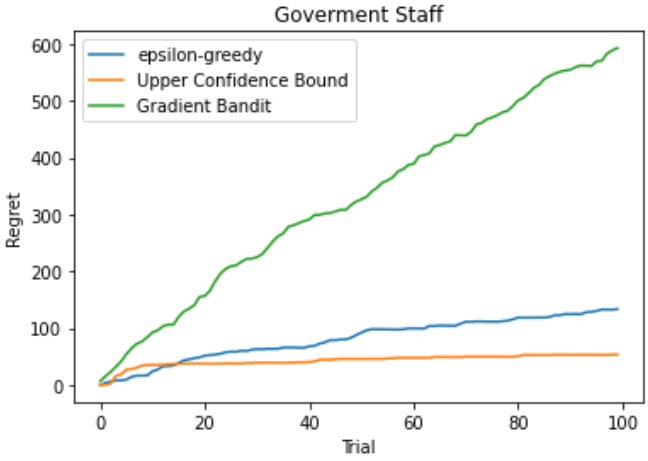
نمودار 1 - متوسط پاداش دریافتی دانشجویان



نمودار 2 - متوسط مقدار پشیمانی دانشجویان



نمودار 3 - متوسط پاداش دریافتی کارمندان



نمودار 4 - متوسط مقدار پشیمانی دریافتی کارمندان

نمودار 5 - متوسط پاداش دریافتی دارندگان مشاغل آزاد

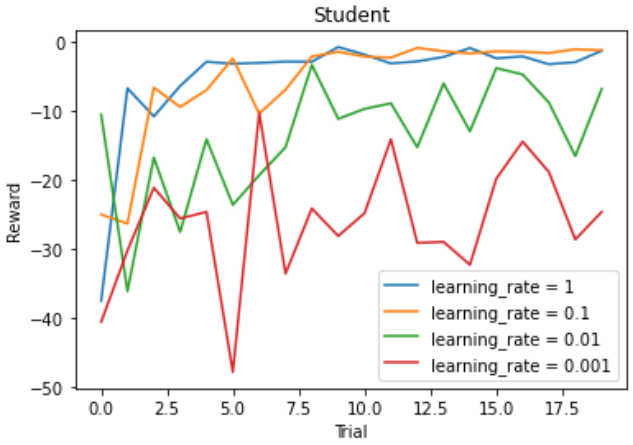
نمودار 6 - متوسط مقدار پشیمانی دارندگان مشاغل آزاد

تحلیل نتایج بخش 4

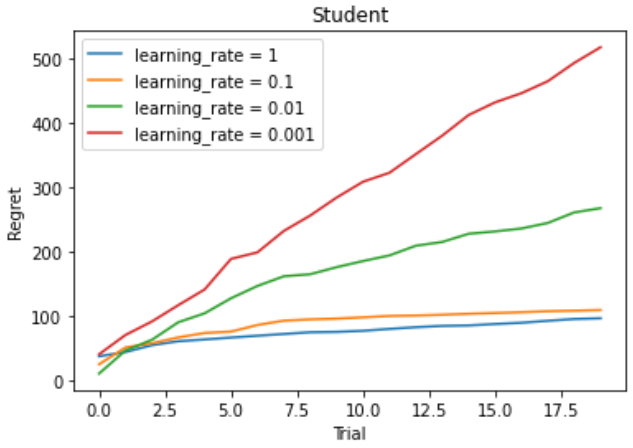
همانطور که مشاهده می‌شود روش Gradient-Bandit بیشترین متوسط پشیمانی را دارد و چرا که این روش با توجه به learning rate = 0.001 برای یادگیری نیاز به زمان زیادی دارد و مشاهده می‌کنیم که اگر تعداد ترایال را خیلی زیاد کنیم این روش به پاداش خوبی همگرا می‌شود.

همچنین روش epsilon-greedy نیز متوسط پشیمانی بیشتری نسبت به روش UCB دارد زیرا با epsilon = 0.2 حتی وقتی از اکشن بهینه مطلع شویم با احتمال 0.2 به سراغ بقیه اکشن ها می‌رویم که این اصلا مطلوب نیست.

بخش 5

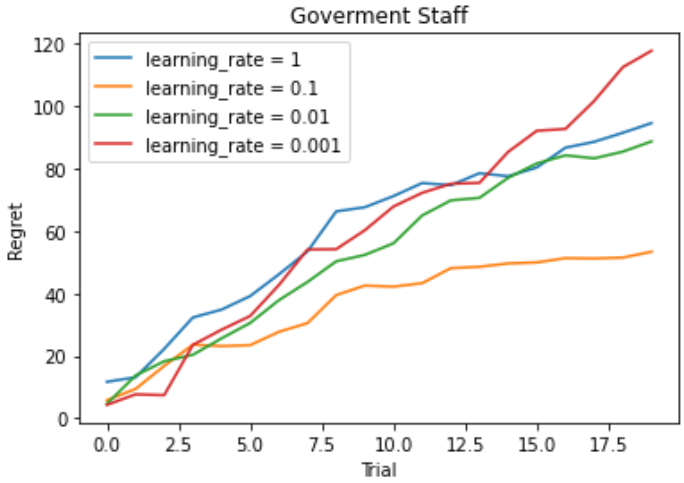


نمودار 7- متوسط پاداش دریافتی دانشجویان به روش گرادیانت



نمودار 8 - متوسط مقدار پشیمانی دانشجویان به روش گرادیانت

نمودار 9 - متوسط پاداش دریافتی کارمندان به روش گرادیانت



نمودار 10 - متوسط مقدار پشیمانی کارمندان به روش گرادیانت

نمودار 11 - متوسط پاداش دریافتی دارندگان مشاغل آزاد به روش گرادیانت

نمودار 12 - متوسط مقدار پشیمانی دانشجویان به روش گرادیانت

تحلیل نتایج بخش 5

با توجه به نتایج گرفته شده برای دانشجویان learning_rate = 1 کمترین متوسط پشیمانی را دارد و برای کارمندان دولتی learning_rate = 0.1 کمترین متوسط پشیمانی را دارد و برای دارندگان مشاغل آزاد learning_rate = 0.01 کمترین متوسط پشیمانی را دارد.

می‌توان این گونه این نتایج را توجیه کرد که هر چه learning_rate ما بیشتر باشد به نوسان های ریوارد بیشتر حساس هستیم و با توجه به نمودار پاداش ها دانش‌آموزان کمترین نوسان و دارندگان مشاغل آزاد بیشترین نوسان را دارند پس به ترتیب لرنینگ ریت بزرگتر و لرنینگ ریت کوچکتر نیاز داریم.

البته باید دقت کرد که با توجه به اینکه تعداد ترایال ها کم است ممکن است با دریافت پاداش های خیلی خوب یا خیلی بد ترتیب این نمودار ها تغییر کند.

در حالت کلی اگر تعداد ترایال به اندازه کافی داشته باشیم با نرخ یادگیری کوچکتر به پاداش مطلوب‌تری همگرا می‌شویم.

منابع