دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

گزارش تمرین شماره یک
درس یادگیری تعاملی
پاییز ۱۴۰۱

نام و نام خانوادگی
سیاوش رزمی
شماره دانشجویی
۸۱۰۱۰۰۳۵۲

فهرست

چکیده

در این تمرین چندین مسأله دنیای واقعی را به شکل multi-bandit مدل سازی می‌کنیم، ســـپس در سؤال ۲ به پیاده‌سازی یک مسأله multi-bandit در پایتون می‌پردازیم.

سوال ۱

هدف سوال:

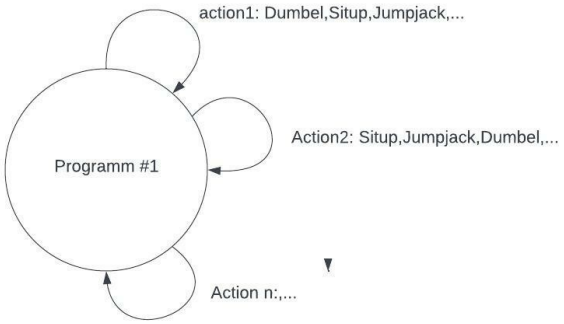
هدف این سؤال مدل کردن تعدادی از مسائل دنیای واقعی با مدل multi-armed bandit است.

۱- در این مسأله ۳ برنامه مختلف برای هر یک از روزهای هفته در نظر گرفته شده که هر کدام شامل تعدادی تمرین ورزشی است برای حل این مسأله هر یک از ۳ برنامه را به شکل یک Bandit در نظر میگیریم به این شکل که هر یک از برنامه‌ها یک context است که ترکیب های مختلف تمرین‌های هر برنامه هر کدام از بازو های این context ها را تشکیل می‌دهند بنابرین هر یک از context ها به تعداد فاکتوریل های تمرین‌ها بازو دارند، میزان پاداش ما مقدار انرژی مصرف شده در هر مرتبه استفاده از آن برنامه مذکور است به طور مثال:

بازو ۱: پشت بازو دمبل،جلو بازو هالتر

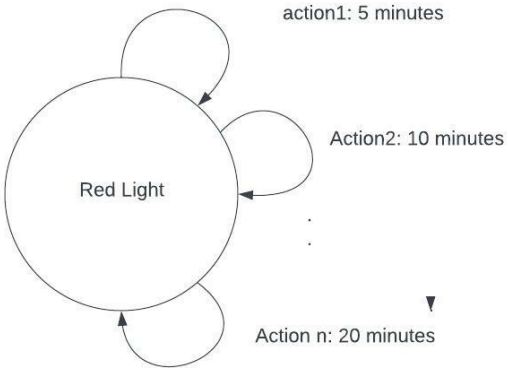
...

بازو n:

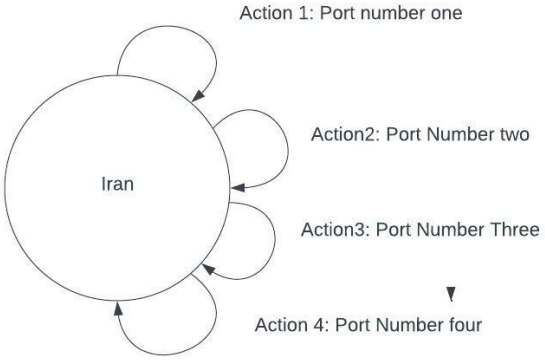


حال برای حل این مسأله میتوان با اجرای بازو ها در هر trial در نهایت با استفاده از سیاست‌های مختلف مانند (greedy، ubc و یا gradient) مسأله را حل کرد.

۲- در این مسأله ما در پشت چراغ قرمز منتظر هستم و در صورتی که چراغ سبز شود پس از ۱۰ دقیقه و در صورتی که به سمت چپ بپیچیم ۳۰ ساعت زمان می‌برد تا به مقصد برسیم، حال برای مدل کردن این مسأله بایستی مقادیر زمان منتظر ماندن پشت چراغ را به شکل گسسته در نظر بگیریم به طور مثال مقادیر گسسته ۵ دقیقه، ۱۰ دقیقه را به شکل بازو در نظر می‌گیریم در این صورت در هر مرتبه که پشت چراغ قرمز ایستاده‌ایم مقادیر مختلف را امتحان می‌کنیم تا میزان بهینه ایستادن پشت چراغ قرمز را با استفاده از سیاست‌های موجود به دست آوریم، میزان پاداش ما قرینه زمان رسیدن به مقصد است.



۳-در این مسأله یک سوئیچ داریم که دارای ۴ درگاه است و بسته هایی که به این مسیریاب می‌رسد به مقاصد کشور های مختلف ارسال می‌شوند، برای مدل کردن این مسأله هر یک از مقاصد بسته هارا یک حالت جداگانه در نظر می‌گیریم و اکشن های هر حالت را درگاهی که می‌توان از آن بسته را به مقصد هدایت کرد در نظر می‌گیریم و میزان پاداش قرینه کل زمان از لحظه دریافت تا زمان دریافت بسته از تصدیق است.

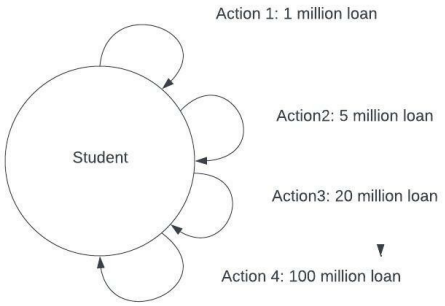


سوال ۲ -

هدف سوال:

در این سؤال به مدل سازی یک مسأله دنیای واقعی و پیاده‌سازی آن در محیط پایتون می‌پردازیم و سعی میکنیم با استفاده از سیاست‌های مختلف آن را حل و نتایج را تحلیل و بررسی کنیم.

۱-در این مسأله سه نوع مشتری داریم که هرکدام با توجه به شرایط خود با احتمال متفاوتی هر یک از مبالغ وام را پرداخت می‌کنند، برای مدل کردن این مسئله هر کدام از مشتریان بانک (دانشجو، کارمند دولتی و صاحب شغل آزاد) به عنوان یک bandit در نظر گرفته می‌شود و اکشن‌ها مبالغ وام پرداختی به آنهاست، میزان پاداش ما برابر است با مبلغ بازگردانده شده توسط مشتری منهای میزان وام پرداخت شده به وی است به این ترتیب در صورتی که مشتری تمام مبلغ وام به همراه کارمزد آن را پرداخت کند بانک به اندازه مابه ازای کارمزد از این وام پاداش دریافت میکند.



-۲

کلاسی به نام Environment تعریف کردیم که با دریافت هر کدام از مشتریان میزان پاداش و اکشن های موجود را با استفاده از دو متد تعریف میکنیم.

در کلاس Agent متد های choose_action و step به شکل abstract تعریف شد تا در هر یک از انواع سیاست‌ها (ucb،greedy و gradient) به شکل متفاوت تعریف می‌شود.

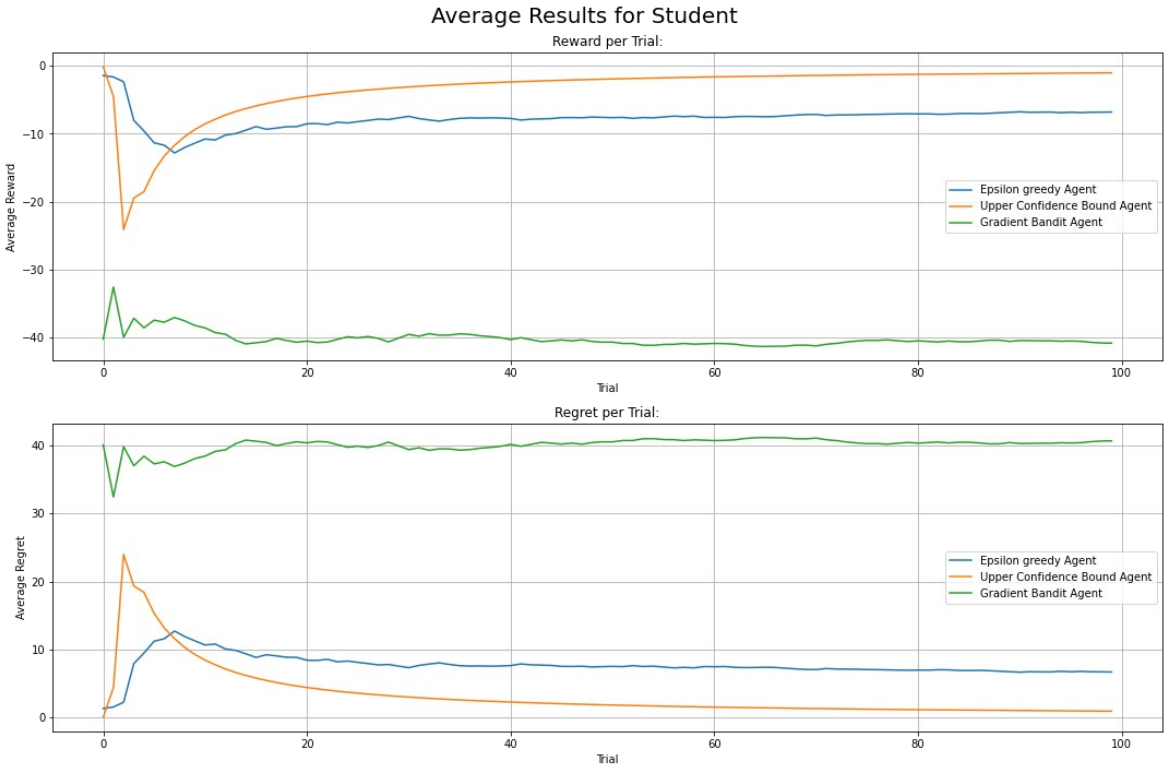
در کلاس epsilon برای پیاده‌سازی سیاست epsilon greedy با تعریف متد choose_action با احتمال epsilon یک عمل رندم و با احتمال یک منهای اپسیلون اکشن با بیشترین utitliy را انتخاب می‌کنیم.

در تابع step اکشن را انتخاب کرده و پاداش آن را بدست آورده، سپس value های اکشن ها را به روز کرده و میزان regret را بدست می آوریم.

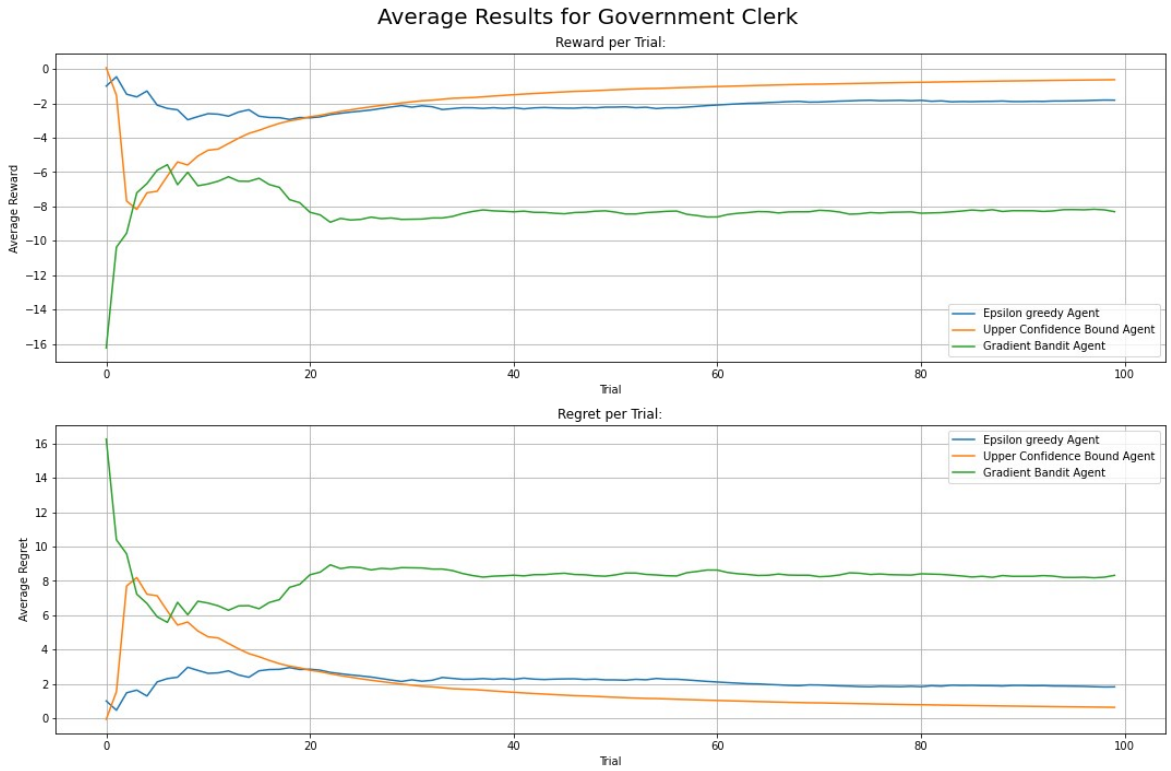
در کلاس gradient ابتدا سیاست هارا به دست آورده و اکشن هارا به نسبت احتمال آن‌ها انتخاب می‌کنیم و در تابع step مقادیر prefrence (h(a)) را به روز میکنیم.

در کلاس ucb در تابع choose_action مقادیر ucb را طبق فرمول به دست می‌آوریم و برای جلوگیری از تقسیم بر صفر زمانی که تعداد باری که اکشن را انتخاب کرده‌ایم صفر باشد مقدار آن را برابر با 1e500 میگذاریم.

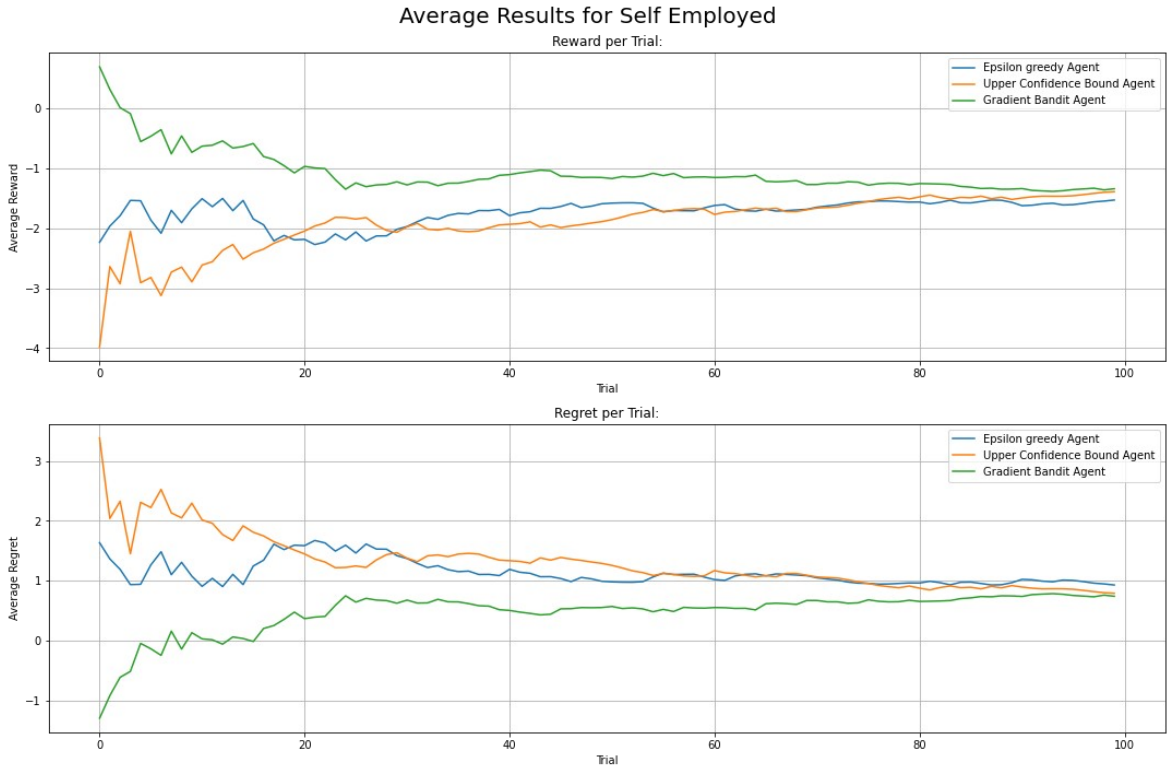
۴- مطابق شرایط گفته شده در سؤال از هر کدام از کلاس‌ها یک object با مشخصات تعریف شده تعریف میکنیم و الگوریتم را ۲۰ مرتبه هرکدام با ۱۰۰ ترایال اجرا میکنیم، که نتایج آن به شکل زیر است:



شکل ۱-۲: نمودار پاداش و پشیمانی برای مشتری دانشجو



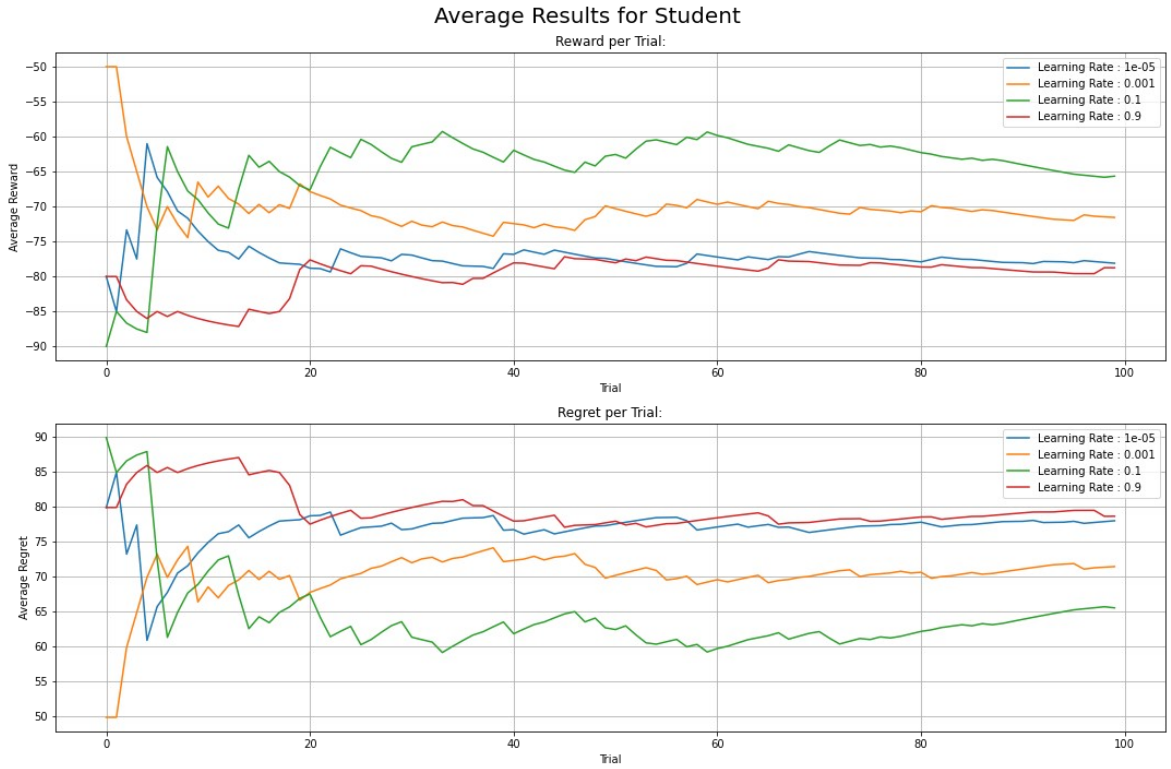
شکل ۲-۲: نمودار پاداش و پشیمانی برای مشتری کارمند دولت



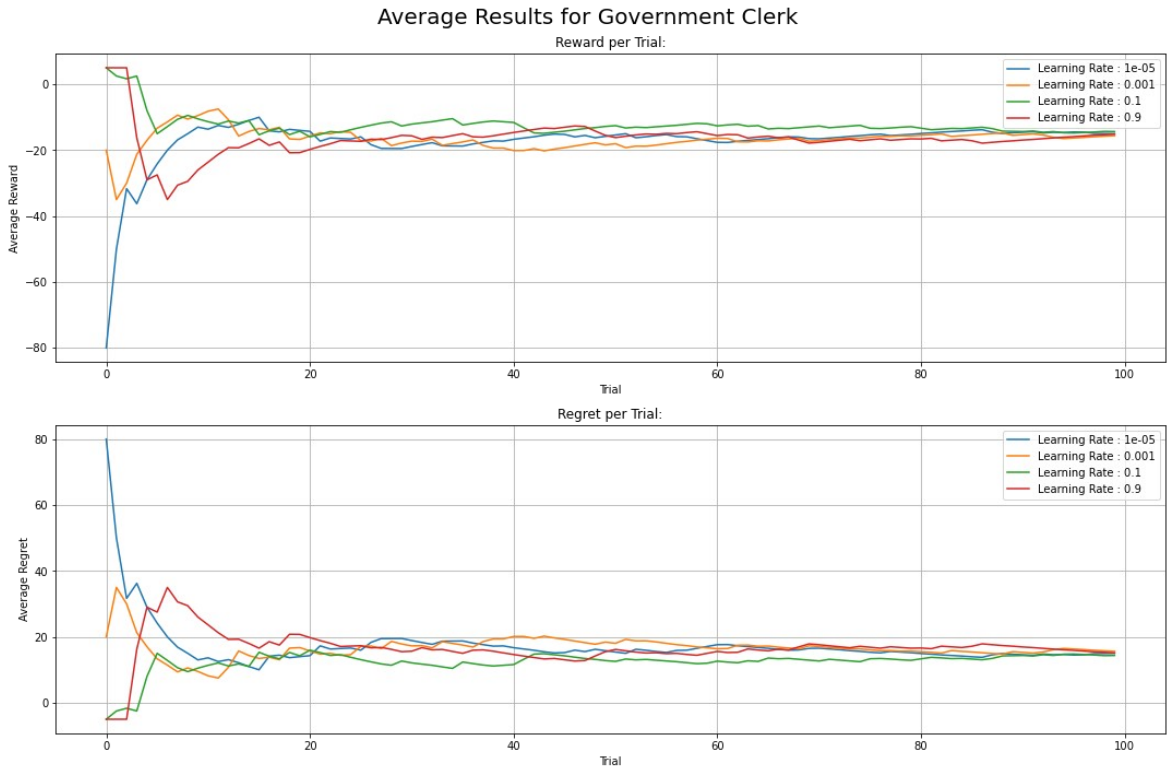
شکل ۳-۲: نمودار پاداش و پشیمانی برای مشتری شاغل آزاد

به نظر می‌رسد که سیاست UCB در مشتری‌های اول و دوم بهتر عمل کرده و بعد از آن سیاست epsilon greedy بهتر عملکرده و سیاست gradient از همه بدتر عمل کرده اما در مشتری شغل آزاد به نظر می‌رسد همه سیاست‌ها در نهایت به نقطه بهینه شده اند.

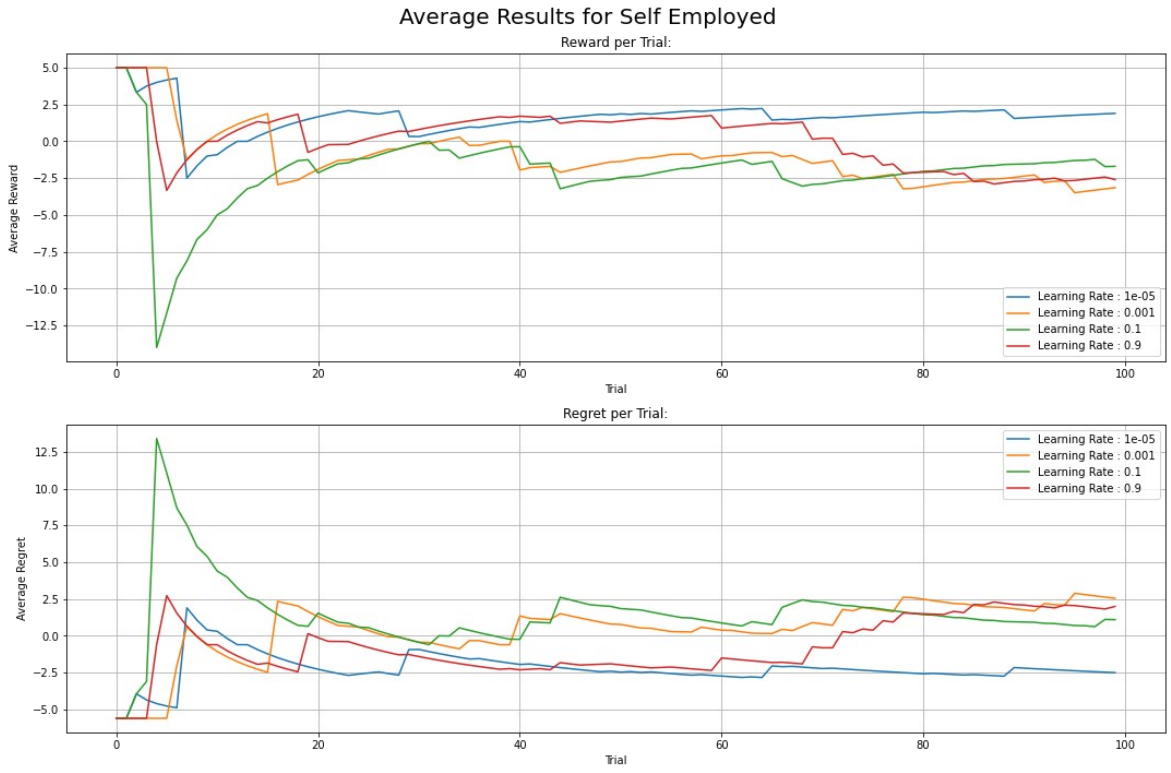
۵ -



شکل ۴-۲: نمودار پاداش و پشیمانی برای مشتری دانشجویی



شکل ۵-۲: نمودار پاداش و پشیمانی برای مشتری کارمند دولت



شکل ۵-۲: نمودار پاداش و پشیمانی برای مشتری شاغل آزاد

به نظر می‌رسد نرخ‌های یادگیری پایینتر عمل‌کرد بهتری داشته‌اند، در مشتری‌های دانشجو و کارمند دولت نرخ ۰.۱ و در شاغل آزاد نرخ ۰.۰۰۰۰۱ عمل‌کرد بهتری داشته است.

نکات مهم و موارد تحویلی

لازم است که به نکات زیر در نوشتن گزارش توجه داشته باشید.

۱. ساختار کلی گزارش که در این فایل به آن اشاره شده باید رعایت شود. در صورت تمایل می‌توانید از latex یا هر نرم افزار دلخواه دیگر برای نوشتن گزارش استفاده کنید، به شرط اینکه ساختار کلی گفته شده رعایت شود. لذا در صورت رعایت نکردن ساختار کلی گزارش بخشی از نمره تمرین کم خواهد شد.
۲. برای تصاویر و جداول موجود در گزارش حتما کپشن قرار داده شود.
۳. **نتایج و تحلیل‌های** شما در روند نمره دهی اهمیت بسیار بالایی دارد، لذا خواهشمندیم کلیه نتایج و تحلیل‌های خواسته شده به صورت کامل و دقیق در گزارش آورده شوند.
۴. در صورت مشاهده شباهت بین گزارش شما و افراد مختلف نمره این سری تمرین برای شما در نظر گرفته نمی‌شود.

موارد تحویلی

۱. برای هر سری از تمرینات، فقط یک فایل با فرمت PDF آماده کنید.
۲. به همراه فایل گزارش، یک پوشه به نام Codes ایجاد کنید و کدها و فایل‌های پیاده‌سازی هر سوال را به صورت تفکیک شده در پوشه‌های جداگانه قرار دهید.
۳. هیچ گونه جدول یا تصویر به صورت جداگانه خارج از گزارش ارسال نشود. مگر اینکه به صورت صریح در تمرین از شما خواسته شده باشد.
۴. در انتها، لطفا برای هر تمرین گزارش و پوشه کدها را به صورت گفته شده، در یک فایل زیپ با فرمت زیر در سامانه یادگیری الکترونیک بارگذاری نمایید.

HW#_LastName_StudentNumber.zip

به طور مثال:

HW1_Mesbah_810111111.zip

منابع

در این بخش منابع (مقالات، سایت‌ها و ...) که در تمرینات و پیاده‌سازی استفاده کرده‌اید را به یک فرمت استاندارد برای مثال فرمت IEEE وارد کنید. برای رفرنس دهی میتوانید از بخش Reference نرم‌افزار Word یا این سایت یا افزونه‌های دیگر استفاده کنید. لازم است منبع مورد استفاده خود را در بخش مربوطه ارجاع دهید.

توجه به این نکته ضرروی است که میزان شباهت کد یا راه حل شما در صورت استفاده از منابع دیگر باید به حد معقولی باشد و کپی کردن از منابع مد نظر ما نیست. در صورت مشاهده کپی از منابع به صورت کامل نمره تمرین یا بخش مورد نظر به شما تعلق نمی‌گیرد.