سؤال یک (سیستم پیشنهاد غذا) (۵۰ نمره)

شما به عنوان یک تیم توسعه‌دهنده، مسئول طراحی سیستمی برای توصیه غذا به کاربران در یک اپلیکیشن هستید. هدف این است که بهترین غذاها را بر اساس سلیقه‌های کاربران پیشنهاد دهید. برای این کار، از الگوریتم‌های یادگیری تقویتی استفاده می‌کنید. هر بازوی به یک نوع غذا اشاره دارد و احتمال پاداش نشان‌دهنده رضایت کاربر از غذای پیشنهادی است.

صورت بندی مسئله:

۱. **تعداد بازوها:**
   - تعداد بازوها برابر است با باقیمانده تقسیم آخرین سه رقم شماره دانشجویی بر ۵ به علاوه ۲.
   - **مثال:** برای شماره دانشجویی ۸۱۰۱۰۲۳۵۳، باقیمانده تقسیم ۳۵۳ بر ۵ برابر ۳ است و بنابراین تعداد بازوها ۵ می‌شود.
۲. **احتمالات پاداش:**
   - N رقم سمت راست شماره دانشجویی خود را انتخاب کرده (N برابر تعداد بازو هاست) و هر کدام را بر ۱۰ نقسیم کرده و برابر احتمال پاداش دادن هر بازو در نظر بگیرید. (در صورت پیروزی ۱ امتیاز مثبت و در غیر ان صورت ۰ دریافت می کنید.) در صورتی که رقمی از شماره دانشجویی شما ۰ است آن را برابر ۵ در نظر بگیرید. برای مثال برای شماره دانشجویی ۸۱۰۱۰۲۳۵۳ احتمال پاداش دادن بازو ها به ترتیب ۰.۳، ۰.۵، ۰.۳، ۰.۲، ۰.۵ میشود.
۳. **پیاده‌سازی الگوریتم‌های:**
   - **Thompson Sampling**

     از توزیع بتا برای هر بازو بر اساس احتمالات تخصیص داده شده استفاده کنید.
   - **Upper Confidence Bound (UCB)**

     پارامتر c=۱ را برای کنترل اکتشاف تعیین کنید.
   - **Value-based Epsilon Greedy**

     مدل تغییر پارامتر epsilon در طول یادگیری را تنظیم کنید.

وظیفه شما:

- **پیاده سازی و مقایسه الگوریتم‌های فوق:**

بر اساس مجموع پاداش‌های دریافتی و سرعت همگرایی به سیاست‌های بهینه، الگوریتم‌ها را مقایسه کنید.

توجه کنید که حتما نمودار آماره های لازم برای مقایسه متوسط پاداش (reward) و پشیمانی (regret) را رسم کنید.

- **بررسی تأثیر تغییر پارامترها:**

تأثیر تغییر پارامترهای هر روش بر نتایج به‌دست‌آمده را تحلیل کنید.

سؤال دو (تبلیغات در یک فروشگاه آنلاین) (۲۰ نمره)

شما به عنوان یک مهندس داده، در حال کار روی یک سیستم توصیه‌گر برای یک فروشگاه آنلاین هستید. فروشگاه شما دو نوع تبلیغ مختلف برای هر محصول دارد: **تبلیغ ۱** و **تبلیغ ۲**

هر کاربری که وارد وب‌سایت می‌شود، یکی از این دو تبلیغ به او نمایش داده می‌شود. اما سیستم شما تنها پس از دو بار نمایش تبلیغ به کاربران، اطلاعاتی در مورد موفقیت تبلیغات (مثلاً کلیک یا خرید) را دریافت می‌کند.

جزئیات مسئله:

۱. **انتخاب اکشن‌ها:** شما دو اکشن در اختیار دارید:
   - اکشن ۱: نمایش تبلیغ ۱
   - اکشن ۲: نمایش تبلیغ ۲
۲. **پاداش:** هر بار که یک کاربر دو تبلیغ متوالی مشاهده می‌کند (مثلاً ابتدا تبلیغ ۱ و سپس تبلیغ ۲)، سیستم شما بعد از تبلیغ دوم پاداشی دریافت می‌کند. این پاداش نشان‌دهنده تعامل موفق کاربر با تبلیغ‌هاست (مثلاً خرید یا کلیک روی تبلیغ). به عبارت دیگر، پاداش تنها بعد از دو بار نمایش تبلیغ به یک کاربر برمی‌گردد.
۳. **هدف:** سیستم شما باید یاد بگیرد که چگونگی نمایش ترتیب تبلیغ ها؛ بهترین نتیجه را در قالب کلیک یا خرید دارد. بنابراین، باید یک الگوریتم بندیت طراحی کنید که به مرور زمان با توجه به اطلاعات پاداش‌ها، ترکیب بهینه‌ای از تبلیغ‌ها را برای نمایش پیدا کند.

وظیفه شما:

۱. **صورت‌بندی مسئله به عنوان یک مسئله N-arm Bandit**

توضیح دهید که چگونه می‌توان این مسئله را به صورت یک مسئله N-arm Bandit فرموله کرد. تعریف کنید که هر اکشن چه چیزی است و چگونه پاداش‌ها به سیستم برمی‌گردند.

۲. **استراتژی اکتشاف و بهره‌برداری (Exploration-Exploitation)** :

سیستم شما باید بین امتحان کردن ترکیب‌های مختلف (اکتشاف) و استفاده از بهترین ترکیب فعلی (بهره‌برداری) تعادل برقرار کند. توضیح دهید چگونه می‌توانید این تعادل را با استفاده از یک روش بندیت مدیریت کنید. این روش را با AB-testing مقایسه کنید.

سؤال سه (موبایل اجتماعی)(۳۰ نمره + ۲۰ نمره امتیازی)

تصور کنید موبایلتان (عامل) در تعامل با شما بایستی سرویس مناسب به شما را (در قالب یک وظیفه تک حالته) یاد بگیرد. پاداش شما به موبایلتان تنک (Sparse) است، یعنی به دلایل مختلف به صورت تصادفی به برخی از اعمال موبایلتان پاداش میدهید.

موبایل شما اعمال دیگر موبایلها (دیگر عامل‌ها) در مقابل کاربرانشان را میبینند (موبایل اجتماعی). اما نکته مهم این است که **موبایل شما نمی‌تواند پاداش سایر موبایلها** را ببیند. روشی برای استفاده از این اطلاعات اضافی برای تسریع یادگیری موبایلتان ارایه دهید.

صورت‌بندی مسئله:

۱. **عامل‌ها(Agents)** :
   - چندین موبایل در محیط وجود دارد که هر کدام سیاست‌های مختلفی را برای انتخاب اکشن‌ها دنبال می‌کنند. موبایل شما یکی از این عامل‌ها است.
۲. **عمل ها(Actions)** :
   - همه عامل ها n عمل دارند.
۳. **پاداش‌ها(Rewards)** :
   - پاداش‌ها به صورت **تنک (sparse)** داده می‌شوند. یعنی به دلایل مختلف هر کاربر به صورت تصادفی به برخی از اعمال موبایلش پاداش میدهد. این احتمال برای هر کاربر متفاوت بوده و متغیر با زمان است.
   - شما نمی‌توانید پاداش سایر عامل‌ها را مشاهده کنید.
٤. **مشاهده عمل های دیگر عامل‌ها:**
   - شما می‌توانید انتخاب‌های دیگر موبایلها را مشاهده کنید.
   - شما **سیاست دیگر عامل‌ها** را از طریق مشاهده اکشن‌های آن‌ها **تخمین زده** و میخواهید سیاست خود را بهبود دهید.
٥. **هدف:**
   - طراحی یک استراتژی که نه تنها بر اساس دریافت پاداش‌های کم‌تعداد خودتان عمل کند، بلکه با توجه به رفتار و اکشن‌های سایر عامل‌ها، بهترین سیاست ممکن را برای رسیدن به بیشترین پاداش کلی را با کمترین تاسف (پشیمانی) توسعه دهد.

وظیفه شما:

۱. **صورت‌بندی مسئله**
   - توضیح دهید که چگونه این مسئله را می‌توان به عنوان یک مسئله یادگیری تقویتی اجتماعی با **پاداش‌های sparse** فرموله می کنید؟ چطور باید تصمیم بگیریم و چگونه از اطلاعات مشاهده‌ای از اکشن‌های دیگر عامل‌ها استفاده می کنید؟

۲. **استراتژی اکتشاف-بهره‌برداری:(Exploration-Exploitation)**

- به دلیل عدم مشاهده مستقیم پاداش‌های دیگر عامل‌ها، عامل باید بین **اکتشاف** رفتارهای جدید و **بهره‌برداری** از سیاست کنونی تعادل برقرار کند. توضیح دهید که چگونه می‌توانند این تعادل را با توجه به sparse بودن پاداش‌ها مدیریت کنید.

۳. **(امتیازی)**

- در ارتباط با دیگر عامل های موجود در محیط و سیاست های آنها فرض معقولی کرده و الگوریتم خود را شبیه سازی کنید. و نتایج مربوط به آن را گزارش کنید.

نکات تمرین

- استفاده از LLM ها در این تمرین مشکلی ندارد. اما در صورت استفاده لطفاً منبع و prompt خود را ذکر نمایید تا تقلب محسوب نشود.
- مهلت ارسال این تمرین تا پایان روز جمعه ۴ آبان ماه خواهد بود.
- انجام این تمرین به‌صورت یک‌نفره است. اما بحث و گفت‌وگو در پیامرسان درس مانعی ندارد.
- دقت کنید که مهم ترین معیار نمره دهی به شما بر اساس گزارش ارائه شده توسط شماست.
- در تمامی سوالات ذکر موارد زیر لازم است:
  - توضیحات پیاده‌سازی و الگوریتم‌های انتخابی
  - شبه کد الگوریتم های پیاده سازی شده و یا الگوریتم های پیشنهادی ارائه شده
  - نحوه تنظیم پارامتر ها
  - نتایج به‌دست‌آمده از اجرای الگوریتم‌ها به همراه بازه اطمینان ۹۵ درصد
  - تحلیل نتایج و مقایسه الگوریتم‌ها

- لطفاً گزارش تمرین را در قالب Pdf و کد تمرین را به صورت Notebook با مشخص نمودن شماره سوال، در سامانه ایلرن بارگذاری نمایید.
- در صورت وجود سؤال و یا ابهام می‌توانید از طریق پیام‌رسان درس با دستیاران آموزشی در ارتباط باشید.