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