تحلیل و بررسی پروژه یادگیری تقویتی

الهه رضاپناه ، زهرا رستمی

9912762858–9912762789

-1

Q-learning: یک الگوریتم off-policy است، به این معنی که از سیاستی مستقل از سیاست فعلی عامل برای به‌روزرسانی ارزش اکشن‌ها استفاده می‌کند.

Q-learning از مقدار Q بیشینه اکشن بعدی (بیشترین مقدار Q برای تمامی اکشن‌های ممکن در حالت بعدی) استفاده می‌کند تا مقدار Q اکشن فعلی را به‌روزرسانی کند. این روش باعث می‌شود که Q-learning تمایل بیشتری به کاوش (exploration) داشته باشد و اغلب به سیاست بهینه سریع‌تر دست یابد.

به همین دلیل سریع ترین راه رسیدن که حرکت از لبه ی دره هست رو انتخاب می‌کند.

SARSA: یک الگوریتم on-policy است، به این معنی که از سیاست فعلی عامل برای به‌روزرسانی ارزش اکشن‌ها استفاده می‌کند.

SARSA از مقدار Q اکشن واقعی که عامل انتخاب کرده است استفاده می‌کند تا مقدار Q اکشن فعلی را به‌روزرسانی کند. این روش باعث می‌شود که SARSA تمایل بیشتری به بهره‌برداری (exploitation) از سیاست فعلی داشته باشد و معمولا رفتار پایدارتر ولی کندتری در یادگیری نشان دهد.

به همین دلیل راه مطمعن تر که بیشترین فاصله را از دره دارد انتخاب می‌کند که با وجود دیرتر رسیدن مطمعن باشد در دره نمی افتد.

تفاوت عملکرد این دو الگوریتم ناشی از نحوه به‌روزرسانی ارزش‌ها و سیاست‌های تصمیم‌گیری آنهاست. در محیط‌های نامعین و پویا، Q-learning به دلیل خاصیت کاوش بیشتر، ممکن است سیاست‌های بهتری پیدا کند ولی در محیط‌های پایدارتر، SARSA ممکن است عملکرد بهتری داشته باشد زیرا از سیاست فعلی پیروی می‌کند و کمتر به تغییرات تصادفی حساس است.

2-

200 episode: در این مرحله، هر دو الگوریتم هنوز در حال کاوش محیط هستند و ممکن است سیاست‌های پایدار و بهینه‌ای پیدا نکرده باشند. Q-learning به دلیل کاوش بیشتر ممکن است سیاست‌های بهتری پیدا کند ولی عملکرد کلی ممکن است ناپایدار باشد.
در تست ما Q-learning همچنان از مسیر درست که در لبه دره هست حرکت کرد و به goal رسید.
اما sarsa راه اشتباهی را به صورت پله ای و عوض کردن سطر در چند مرحله رفت و با وجود رسیدن به goal از سیاست های تعریف شده اش استفاده نکرد.

250 episode: در این مرحله، الگوریتم‌ها شروع به تثبیت سیاست‌های خود می‌کنند. SARSA به دلیل پیروی از سیاست فعلی ممکن است سیاست‌های پایدارتری پیدا کند ولی Q-learning همچنان در حال کاوش و بهبود سیاست‌ها است.
در تست ما Q-learning مجددا همچنان از مسیر درست که در لبه دره هست حرکت کرد و به goal رسید.
درحالی که sarsa همچنان راه اشتباهی را به صورت پله ای و عوض کردن سطر در چند مرحله رفت و با وجود رسیدن به goal از سیاست های تعریف شده اش استفاده نکرد.

300 episode: در این مرحله، هر دو الگوریتم به سیاست‌های پایدار و بهینه نزدیک شده‌اند. Q-learning به دلیل کاوش بیشتر ممکن است سیاست‌های بهتری پیدا کند ولی SARSA همچنان سیاست‌های پایدارتری خواهد داشت.
در تست ما Q-learning مجددا هم از مسیر درست که در لبه دره هست حرکت کرد و به goal رسید.
Sarsa با این مقدار اپیزود مسیری صاف و بدون تعویض سطر رفت ولی به جای دور ترین مسیر از دره از وسط حرکت کرد.

به طور کلی، با افزایش تعداد اپیزودها، هر دو الگوریتم عملکرد بهتری نشان می‌دهند ولی Q-learning به دلیل خاصیت کاوش بیشتر ممکن است سیاست‌های بهینه‌تری پیدا کند در حالی که SARSA سیاست‌های پایدارتر ولی کندتری خواهد داشت.

در تست های ما Q-learning در هر سه حالت همچنان درست کار کرد ولی Sarsa مسیری به صورت پله ای و عجیب را طی کرد که نشان دهنده learn نشدن به صورت درست هست در نهایت در 300 مسیر بهتری نزدیک به مسیر درستش پیدا کرد ولی بازهم کامل learn نشده بود (گیف های حرکات به همراه نمودارها در هر تعداد اپیزود در پوشه به همان نام قرار دارد)