
# درس یادگیری ماشین

# گزارش پروژه سوم

## تهیه و تنظیم

سید عماد حلمی (۹۶۳۰۸۲۵۰۸)

## استاد درس

**دکتر میرمحسن پدرام**

**زمستان ۱۳۹۶**

**هدف پروژه:**

در این پروژه هدف آشنایی با دو الگوریتم معروف در یادگیری تقویتی به نام های Qlearning و SARSA است.

**پیاده‌سازی:**

دو فایل در پوشه‌ی کد ها با پسوند پایتون وجود دارند که هرکدام جداگانه یکی از این الگوریتم ها را پیاده‌سازی کرده‌اند. برای پیاده‌سازی محیط نیز از یک فایل csv با نام map.csv استفاده شده است که در آن یک ماتریس با ابعاد ۳۶*۳۶ را ساخته ایم.که هر خانه نشان دهنده‌ی میزان reward از خانه‌ی متعلق به سطر به خانه‌ی متعلق به ستون است. برای سادگی، اگر از خانه‌ای به خانه‌ای راه وجود نداشته‌باشد، مقدار آن خانه از ماتریس ۲- گذاشته می‌شود.

**پاسخ سؤال‌های پروژه:**

**الف) یک اپیزود چگونه تعریف می‌شود؟**

یک اپیزود درواقع یک مسیر از خانه‌ی شروع(به صورت تصادفی انتخاب می‌شود) به خانه‌ی هدف است.

**ب) جدول Q را چگونه تعریف می‌کنید؟**

به دلیل آنکه این الگوریتم‌ها هر دو الگوریتم هایی بازگشتی هستند، یک مقدار کوچک(محدود) که به نام «optimistic initial conditions» نیز شناخته می‌شود، برای نایل شدن به جواب نهایی کافی است. در این دو کد این ماتریس ها از ابتدا صفر گذاشته شده‌اند.

**ج و د) عامل را با دو الگوریتم گفته شده آموزش دهید و برای انتخاب عمل از روش e-Greedy استفاده کنید.**

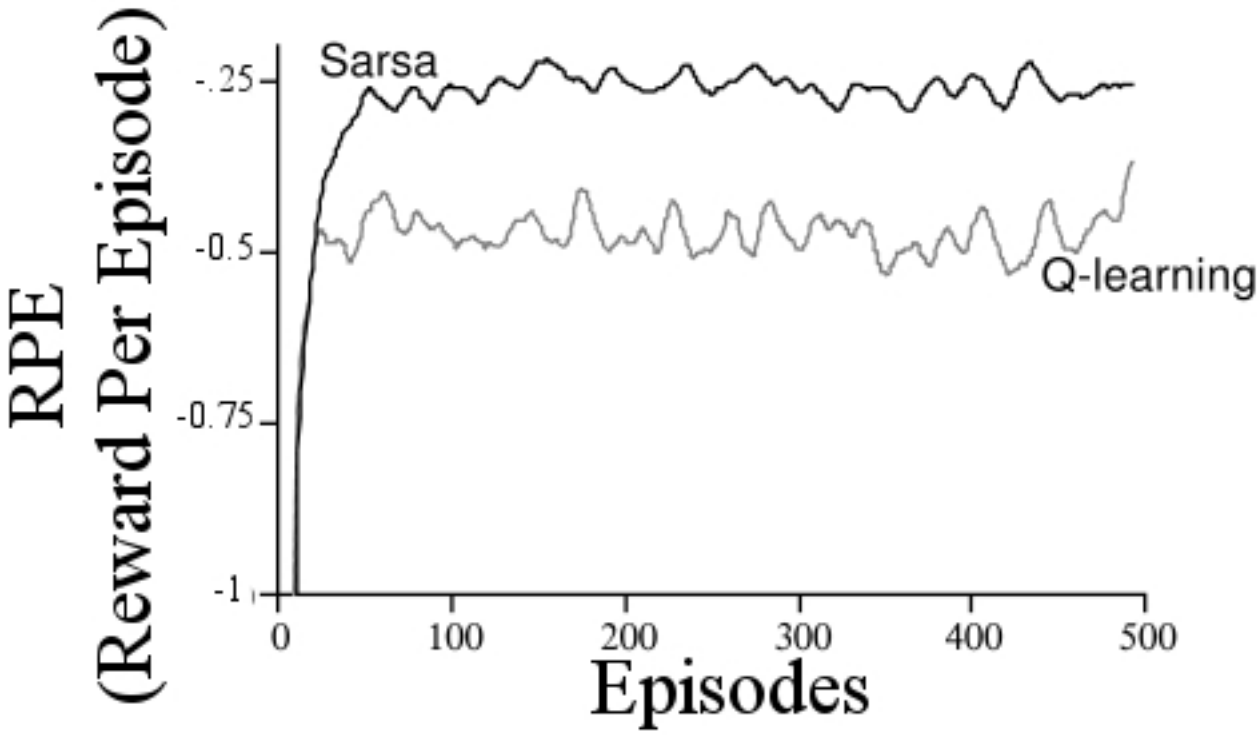
دو فایل با پسوند csv با نام های qlearning_q.csv و sarsa_q.csv در پوشه‌ی فایل موجودند که هر کدام ماتریس نهایی Q حاصل از اجرای ۱۰۰۰۰ اپیزود از الگوریتم های فوق را با حالت‌های اولیه‌ی تصادفی را ذخیره کرده‌اند. روش e-Greedy نیز بدین صورت است که با احتمال e بدون توجه به هیچ مقداری و یادگیری ای، یکی از action های موجود را به طور اتفاقی انتخاب می‌کند و با احتمال مکمل e از بین action های مجود بهترین action را انتخاب می‌کند. بهترین action نیز آن است که مقدار Q آن در جدول بیشترین باشد. طبیعی است در صورت یکسان بودن یک یا چند action یکی از آن‌ها به صورت تصادفی انتخاب می‌شود.

**ه) مقادیر پارامتر های یادگیری را چگونه انتخاب می‌کنید؟ چرا؟**

با توجه به اینکه خانه‌های ممنوعه همگی reward منفی دارند، بهتر است مقدار gamma که در هر دو الگوریتم استفاده شده است کم باشد. میدانیم که هرچه نرخ تخفیف کمتر باشد الگوریتم فرصت طلبانه تر به گزینه ها نگاه می‌کند و در اینجا چون برای نقاط ممنوعه میزان پاداش منفی است و برای سایر نقاط به جز منتهی شونده های به هدف این مقدار صفر است پس با مقایسه‌های زیادی سر و کار نداریم و تمایز همواره معلوم و مشهود خواهد بود. برای این کار gamma که همان نرخ تخفیف است مساوی ۰.۲ در نظر گرفته شده است. و در الگوریتم SARSA نیز علاوه بر نرخ تخفیف با نرخ یادگیری مواجه هستیم که این نرخ اگر کم باشد یعنی نیازی به یادگیری چیزی ندارد و اگر خیلی زیاد باشد(یعنی ۱) به این معناست که آخرین اطلاعات را در نظر بگیرد. برای این مثال و شرایط موجود مقدار ۰.۸ برای نرخ یادگیری یا همان alpha در نظرگرفته شده است.

**و و ز) نمودار های مربوطه را برای هر دو الگوریتم رسم کرده و حاصل را برای این مسأله خاص تحلیل کنید.**

با توجه به نمودار زیر که نمودار Reward تجمعی در هر episode است که برای در حدود ۵۰۰ اپیزود اجرا شده است میبینیم که الگوریتم SARSA از حیث این مقدار نتایج بهتری می‌دهد. دلیل آن نیز این است که این الگوریتم در اصطلاح تلاش در یادگیری safe path یعنی مسیری که از مانع ها دور تر باشد را دارد.




*نمودار مقایسه الگوریتم های SARSA و QLearning*