به نام خدا

پروژه درس نوروساینس

استاد درس: دکتر آقاجانی کربلائی

ارائه دهنده:

شایان رضائی

شماره دانشجویی: 400108558

ساخت محیط و تنظیمات اولیه

ابتدا محیطFrozenLake را با استفاده از کتابخانه gymnasium ایجاد کردیم. در این مرحله، از محیط پیش‌فرض FrozenLake-v1 استفاده کردیم و گزینه is_slippery را برابر True قرار دادیم تا شرایط غیرقطعی ایجاد شود(در ادامه شرایط قطعی را نیز پیاده سازی خواهیم کرد.)

جدول حاصل شده در این محیط شامل دو بخش میشود:

بخش states که فضای حالت ما را تشکیل میدهد و آن را یک جدول چهار در چهار قرار دادیم (هر خانه میتواند چهار حالت عادی، سوراخ، شروع و پایان را داشته باشد). به همراه بخش actions که شامل چهار تصمیم چپ، راست، بالا و پایین میباشد.

مرحله ۲: آموزش عامل باQ-Learning

الگوریتم Q-Learning با پارامترهای زیر آموزش داده شد:

- نرخ یادگیری0.8 :(alpha)
- ضریب تخفیف0.99 :(gamma)
- مقدار اولیه اپسیلون1.0 :(epsilon)
- نرخ کاهش اپسیلون: 0.995
- حداقل اپسیلون: 0.01
- تعداد اپیزودهای آموزشی: 10000
- حداکثر گام در هر اپیزود: 100

در طی آموزش، عامل با استفاده از سیاست epsilon-greedy اقدام به کاوش و یادگیری پاداش‌ها نمود. با گذر زمان و کاهش اپسیلون، عامل به سمت بهره‌برداری بیشتر از سیاست بهینه حرکت کرد.

مرحله ۳: ارزیابی عملکرد عامل

پس از آموزش، عامل آموزش‌دیده در 100 اپیزود آزمایشی ارزیابی شد. در این مرحله فقط از بهترین عمل بر اساس Q-Table استفاده گردید (بدون کاوش تصادفی).

نتیجه ارزیابی:

- درصد موفقیت در رسیدن به هدف **77 درصد**

این مقدار نشان‌دهنده عملکرد قابل قبول عامل در شرایط غیرقطعی محیط است.

مرحله ۴: سیاست نهایی یادگرفته‌شده

هر خانه در شبکه محیط با یکی از چهار عمل ممکن (چپ، پایین، راست، بالا) مقداردهی شده است. سیاست یادگرفته‌شده به تفکیک حالات به صورت زیر گزارش می‌شود:

```
[ ['←'  '↑'  '↑'  '↑']
  ['←'  '←'  '←'  '←']
  ['←'  '←'  '↓'  '↑']
  ['←'  '→'  '↓'  '←'] ]
```

در این جدول، در هر خانه سیاست بهینه نمایش داده شده و نشان میدهد که محرک در هر وضعیت چه تصمیمی را اتخاذ میکند.

هم چنینی جدول Q-Table نهایی، که پاداش‌های یادگرفته‌شده برای هر حالت و هر عمل را نشان می‌دهد، به صورت زیر است ، این جدول نشان میدهد که ارزش هر action در هر وضعیت چقدر میشود، بعضی خانه ها که کل سطر صفر است، نشان میدهند که آن خانه یا حالت سوراخ است یا حالت هدف است که محرک در این نقاط حرکتی نخواهد کرد:

	←	↓	→	↑
S 0:	0.616	0.394	0.475	0.496
S 1:	0.000	0.148	0.357	0.612
S 2:	0.160	0.116	0.140	0.610
S 3:	0.240	0.166	0.150	0.380
S 4:	0.649	0.128	0.589	0.561
S 5:	0.000	0.000	0.000	0.000
S 6:	0.576	0.000	0.000	0.000
S 7:	0.000	0.000	0.000	0.000
S 8:	0.133	0.142	0.075	0.580
S 9:	0.119	0.600	0.011	0.242
S10:	0.887	0.003	0.001	0.005
S11:	0.000	0.000	0.000	0.000
S12:	0.000	0.000	0.000	0.000
S13:	0.028	0.130	0.955	0.037
S14:	0.487	0.998	0.607	0.356
S15:	0.000	0.000	0.000	0.000

1. پاداش نهایی هر خانه در برای هر تصمیم

تفاوت شرایط قطعی و تصادفی

در آزمایش دوم، الگوریتم Q-Learning در شرایط قطعی اجرا گردید؛ به این صورت که ویژگی لغزندگی غیرفعال شد .(is_slippery=False) در این حالت، عامل دقیقاً به همان جهتی که انتخاب می‌کند حرکت می‌کند و محیط فاقد رفتار تصادفی است. نتایج حاصل از آموزش نشان‌دهنده یادگیری کامل و مؤثر عامل می‌باشد؛ به‌طوری‌که در مرحله ارزیابی، عامل در ۱۰۰ درصد اپیزودهای تست موفق به رسیدن به هدف شد .جدول سیاست نهایی نیز نشان می‌دهد که عامل به‌طور کامل مسیرهای بهینه را شناسایی کرده

و از چاله‌ها اجتناب می‌کند. همچنین، مقادیر Q-Table به‌خوبی همگرایی یافته‌اند و بیشترین مقادیر مربوط به مسیرهایی هستند که عامل را به‌سوی هدف هدایت می‌کنند. این نتایج تأییدی بر سادگی یادگیری در محیط‌های قطعی در مقایسه با شرایط غیرقطعی هستند.

تاثیر پارامتر های مختلف بر یادگیری و نرخ موفقیت

- بررسی تئوریک پارامترها درQ-Learning

در الگوریتمQ-Learning ، سه پارامتر کلیدی نقش بسیار مهمی در فرایند یادگیری و همگرایی عامل دارند.

اولین پارامتر، نرخ یادگیری (alpha) است که تعیین می‌کند عامل تا چه حد به تجربه جدید نسبت به اطلاعات قبلی خود اعتماد کند. اگر alpha خیلی پایین باشد، عامل به آرامی یاد می‌گیرد و ممکن است برای همگرایی به اپیزودهای بسیار زیادی نیاز داشته باشد. در مقابل، مقدار بسیار بالا برای alpha باعث ناپایداری در یادگیری می‌شود و ممکن است عامل نتواند میان نویز و اطلاعات مفید تفاوت قائل شود. پارامتر دوم، ضریب تخفیف(gamma) ، اهمیت پاداش‌های آینده را نسبت به پاداش فعلی تعیین می‌کند. مقدار پایین gamma به این معنی است که عامل تنها به پاداش‌های فوری اهمیت می‌دهد، در حالی که مقدار بالا باعث می‌شود عامل به دنبال کسب پاداش بلندمدت باشد. در محیط‌هایی که رسیدن به هدف نیازمند چندین گام متوالی است، gamma بالا معمولاً عملکرد بهتری دارد.

سومین پارامتر، نرخ کاهش اپسیلون(epsilon_decay) ، نقش مهمی در توازن بین اکتشاف (exploration) و بهره‌برداری (exploitation) دارد. اپسیلون اولیه بالا

موجب می‌شود عامل در ابتدای یادگیری مسیرهای مختلف را کشف کند و با کاهش تدریجی آن، عامل به مرور تمرکز خود را روی سیاست بهینه متمرکز می‌کند. اگر نرخ کاهش اپسیلون خیلی سریع باشد، عامل زودتر از موعد وارد بهره‌برداری می‌شود و ممکن است فرصت شناسایی مسیرهای بهتر را از دست بدهد.

- تحلیل تجربی عملکرد عامل بر اساس نمودارها

نتایج تجربی حاصل از آزمایش‌های انجام‌شده نشان می‌دهد که تأثیر هر یک از پارامترهای ذکرشده به نحوی قابل توجه بوده است. در رابطه با نرخ یادگیری(alpha) ، مشاهده شد که عامل در دو مقدار پایین (۰,۱) و بالا (۰,۸) توانسته به موفقیت کامل (٪۱۰۰) دست یابد، در حالی که در مقدار میانی (۰,۴) عملکرد به‌شدت افت کرده است. این مسئله می‌تواند نشان‌دهنده حساسیت خاص الگوریتم به مقدار متوسط alpha باشد که شاید در شرایط خاصی باعث نوسان یا همگرایی ناقص شود. در مورد ضریب تخفیف (gamma)، عامل در هر سه مقدار انتخاب‌شده عملکرد ثابتی از خود نشان داد و موفقیت کامل داشت. این ثبات می‌تواند ناشی از ساختار ساده و قطعی محیط باشد که موجب شده وابستگی به پاداش‌های آینده اهمیت کمتری پیدا کند. در نهایت، بررسی نرخ کاهش اپسیلون نشان داد که مقدار بالا (۰,۹۹۸) منجر به موفقیت کامل شده است، در حالی که مقادیر پایین‌تر باعث شکست کامل عامل شده‌اند. این نتیجه به‌خوبی نشان می‌دهد که حفظ فرآیند اکتشاف در مراحل اولیه آموزش، شرط لازم برای یادگیری موفق سیاست بهینه در این محیط بوده است.

نتیجه گیری

در این پروژه، الگوریتم Q-Learning با هدف آموزش یک عامل برای یافتن مسیر بهینه در محیط FrozenLake پیاده‌سازی و ارزیابی شد. ابتدا الگوریتم در محیط غیرقطعی اجرا شد و عامل با درصد موفقیت قابل قبول توانست سیاست مناسبی برای رسیدن به هدف یاد بگیرد. سپس با اجرای همان الگوریتم در محیط قطعی، عامل عملکرد بسیار بهتری از خود نشان داد و موفق به یادگیری کامل سیاست بهینه گردید. در ادامه، تأثیر سه پارامتر کلیدی الگوریتم شامل نرخ یادگیری، ضریب تخفیف و نرخ کاهش اپسیلون مورد بررسی قرار گرفت. نتایج نشان داد که تنظیم درست این پارامترها نقش بسیار مهمی در کیفیت یادگیری و موفقیت عامل دارد، به‌ویژه نرخ کاهش اپسیلون که تأثیر مستقیمی بر توانایی عامل در اکتشاف مسیرهای مطلوب دارد. به‌طور کلی، این مطالعه نشان داد که ترکیب درستی از ساختار محیط، طراحی مناسب الگوریتم و تنظیم دقیق پارامترها می‌تواند منجر به یادگیری موفق سیاست‌های بهینه در مسائل تقویتی شود.