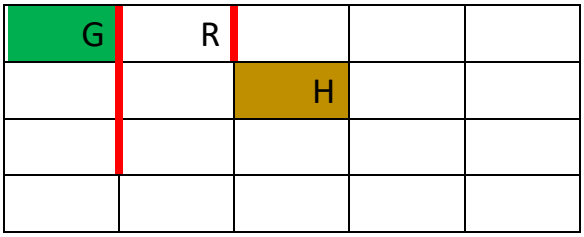
هوش مصنوعی	فصل هفتم: یادگیری تقویتی	تمرین پنجم
پروژه عملی	الگوریتم Q-Learning	تاریخ: ۲۹ آذر ۱۳۹۹

محیط بازی زیر را در نظر بگیرید:



در این محیط ۲۵ مکان مجزا وجود داشته که عامل می‌تواند در هر لحظه در یکی از این مکان‌ها قرار بگیرد که در شروع هر اپیزود به طور تصادفی محل قرارگیری آن مشخص می‌شود. در هر اپیزود، نقطه‌ی هدف مکان G می‌باشد. زمانی که عامل به هدف برسد، پاداش ۱۰ را دریافت خواهد کرد. مکانی که با برچسب H مشخص شده است، محل قرارگیری حفره‌ای را نشان می‌دهد که در صورت برخورد عامل با آن، دچار "آسیب‌دیدگی" می‌شود. اگر عامل مورد نظر قبلاً دچار "آسیب‌دیدگی" شده باشد، پاداش ۱۰- را دریافت خواهد کرد. عامل مورد نظر با قرارگیری در محلی که با برچسب R مشخص شده، تعمیر می‌شود.

در این مسئله، هر حالت به شکل (X,Y,D) نشان داده می‌شود که در آن، X و Y نمایانگر مختصات محل قرارگیری عامل، و D یک متغیر دودویی بوده و اگر مقدار صفر داشته باشد یعنی عامل ما، سالم و بدون آسیب‌دیدگی و اگر مقدار یک داشته باشد یعنی در حال حاضر عامل دچار آسیب‌دیدگی شده است.

الف) با توضیحات فوق، فضای حالت این مسئله، چند حالت مجزا خواهد داشت؟

کنش‌های مجاز برای عامل حرکت به چهار جهت چپ، راست، بالا و پایین در نظر گرفته شده است. به جز کنش"بالا"، سایر کنش‌های عامل به شکل قطعی عمل می‌کنند. در واقع، عامل با انجام کنش "بالا" در هر حالت، با احتمال ۰,۶ به سمت بالا و با احتمال ۰,۴ در جهت عکس حرکت خواهد کرد. اگر عامل با انجام کنشی با دیوارهایی که با رنگ قرمز مشخص شده برخورد کند و یا به خارج از صفحه هدایت شود، پاداش ۱- را دریافت کرده و در همان حالت باقی خواهد ماند. در غیر این صورت، انجام هر کنش در هر حالت پاداشی را به دنبال نخواهد داشت.

ب) الگوریتم Q-Learning با استفاده از سیاست $\varepsilon - greedy$ و در حالت‌های زیر اجرا نمایید:

۱- تعداد اپیزود = ۱۰۰۰، نرخ یادگیری (α) = ۰,۱، نرخ تخفیف (γ) = ۰,۵، نرخ اکتشاف (ε) = ۰,۲ و مقدار Q اولیه برای تمامی حالت‌ها و کنش‌ها برابر صفر در نظر بگیرید.

۲- تعداد اپیزود = ۱۰۰۰، نرخ یادگیری (α) = ۰,۱، نرخ تخفیف (γ) = ۰,۵، نرخ اکتشاف (ε) = ۰,۲ و مقدار Q اولیه برای تمامی حالت‌ها و کنش‌ها برابر ۲۰ در نظر بگیرید.

۳- تعداد اپیزود = ۱۰۰۰، نرخ یادگیری (α) = ۰,۱، نرخ تخفیف (γ) = ۰,۵، نرخ اکتشاف (ε) = ۰ و مقدار Q اولیه برای تمامی حالت‌ها و کنش‌ها برابر صفر در نظر بگیرید.

۴- تعداد اپیزود = ۱۰۰۰، نرخ یادگیری (α) = ۰,۱، نرخ تخفیف (γ) = ۰,۱، نرخ اکتشاف (ε) = ۰,۲ و مقدار Q اولیه برای تمامی حالت‌ها و کنش‌ها برابر صفر در نظر بگیرید.

۵- تعداد اپیزود = ۱۰۰۰، نرخ یادگیری (α) = ۰,۹، نرخ تخفیف (γ) = ۰,۵، نرخ اکتشاف (ε) = ۰,۲ و مقدار Q اولیه برای تمامی حالت‌ها و کنش‌ها برابر صفر در نظر بگیرید.

- نتایج به دست آمده از اجرای مورد اول را با مورد دوم مقایسه کنید.
- نتایج به دست آمده از اجرای مورد اول را با مورد سوم مقایسه کرده و توضیح دهید که چه اتفاقی میافتد اگر عامل همواره در هر حالت کنشی را انتخاب کند که بیشترین مقدار Q(s,a) را در پی داشته باشد؟
- نتایج به دست آمده از اجرای مورد اول را با مورد چهارم مقایسه کرده و بیان کنید کدامیک بهتر از دیگری عمل می‌کند.
- نتایج به دست آمده از اجرای مورد اول را با مورد پنجم مقایسه کرده و بیان کنید کدامیک بهتر از دیگری عمل میکند. چرا تغییر در پارامتر نرخ یادگیری چنین نتیجه‌ای را در پی دارد؟

نکات:

- برای مقایسه‌ی نتایج در هر یک از موارد خواسته شده از نمودار پاداش-اپیزود استفاده نماید (محور افقی نشان دهنده شماره اپیزود و محور عمودی نشان دهنده میانگین پاداش به دست آمده در آن اپیزود است).
- مسیری که عامل را از حالت اولیه به حالت پایان می‌رساند، به همراه مقدار ارزش کنش‌هایی که در بین این مسیر انجام می‌شود (Q(s,a)) در قالب یک جدول در گزارش خود، نشان دهید.