HW5 کتبی

MDP

امیرعلی شهریاری
۸۱۰۱۰۰۱۷۳

۱)

$$V_{i+1}(s) = \max_a \left(\sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V_i(s')\right) \right)$$

s	(۱,۱)	(۱,۲)	(۱,۳)	(۲,۱)	(۲,۲)	(۲,۳)
V_0	۰	۰	-۵	۰	۰	۵
V_1	۰	۰	-۵	۰	۰/۸×۵ = ۴	۵
V_2	۰	۰/۹×۰/۸×۴ + ۰/۱×-۵ = ۲/۴۸	-۵	۰/۸×۰/۹ ×۴ = ۲/۸۸	۰/۸×۵ + ۰/۱×۴ = ۴/۴	۵

۲)

s	(۱,۱)	(۱,۲)	(۱,۳)	(۲,۱)	(۲,۲)	(۲,۳)
$\pi^*(s)$	بالا	چپ	-	بالا	بالا	-

۳)

۱) (۱,۱) - (۱,۲) - (۱,۳)

۱۱) (۱,۱) - (۱,۲) - (۲,۳)

۱۱۱) (۱,۱) - (۲,۱) - (۲,۲) - (۲,۳)

برای محاسبه تخمین‌های Monte Carlo میانگین Reward های دریافتی را در مسیرهایی که از حالت بیان‌شده عبور کرده است را میانگین می‌گیریم.

$$V(۱,۱) = \frac{-۵+۵+۵}{۳} = \boxed{۱/۶۶}$$

$$V(۲,۲) = \frac{۵+۵}{۲} = \boxed{۵}$$

امیرعلی شهریاری (۸۱۰۱۰۰۱۷۳)

۴)

شکل کلی TD-learning : $V(S) = V(S) + \alpha(R + \gamma V(S') - V(S))$

پس از مرحله نخست

همه آپدیت ها جز این مورد صفر خواهد بود

$\therefore V(1,2) = 0 + 0.1(-5 + (0.9 \times 0) - 0) = -0.5$

پس از مرحله دوم به این شکل خواهد بود:

$$\begin{cases} V(1,1) = 0 + 0.1(0 + (0.9 \times -0.5) - 0) = -0.045 \\ V(1,2) = -0.5 + 0.1(0 + (0.9 \times 0) + 0.5) = -0.45 \\ V(2,2) = 0 + 0.1(5 + (0.9 \times 0) - 0) = 0.5 \end{cases}$$

DQN

۱) DQN به احتمال زیاد Deep Reinforcement learning که همانطور که از نامش پیداست، ترکیب کننده یادگیری عمیق و RL است. RL با آزمون و خطا تصمیم گرفته و بخش Deep آن یادگیری عمیق در حال آن و تقریبا از داده های ورودی بدون ساختار و مهندسی فیچر ها فضاهای حالت تصمیم گیری کند.

ورودی های بسیار بزرگ می گیرند و تصمیمات بهینه سازی آنرا انجام می دهد در زمینه های NLP ربات تیک، بازی های ویدیویی و بینایی ماشین استفاده می شوند.

در مقابل از مسئله های تصمیم گیری، حالات S از MDP دارای ابعاد بالا هستند مثل تصاویر دوربین، که RL های سنتی در حل آنها مشکل دارند ولی DQN ها این چنین MDP هایی را حل می کنند [Cal5].

الگوریتم های مختلف و مطالعاتی در شبکه های DQN همگرایی، یادگیری مشروط به هدف و یادگیری چندعاملی از جمله این پژوهش ها هستند.