1- یکی از دانشجویان درس RL که برای تعطیلات نوروزی به استرالیا سفر کرده, هواپیما او دیر وقت به زمین می‌نشیند. در این منطقه ساحلی وجود دارد که در طول روز یا همواره طوفانی و یا همواره آرام است. او متوجه می شود دانشجو سال گذشته درس که اکنون به این کشور مهاجرت کرده وضعیت جوی ساحل را به صورت یک مدل Markov مدل کرده. اگر هنگام فرود هواپیما (S0) مشاهده کند که ساحل آرام است، چقدر احتمال دارد او بتواند پس فردا (S2) به ساحل سر بزند؟ (در جدول زیر c معادل آرام یا Calm و T معادل طوفانی یا turbulent در نظر گرفته شده)(۱۰ نمره)

S_{t-1}	S_t	$P(S_t \mid S_{t-1})$
C	C	0.9
C	T	0.1
T	C	0.5
T	T	0.5

جواب:

از آنجایی که در مدل مارکوف برای پیش بینی هر استیت تنها به اطلاعات state پیشین نیازمندیم, باید از روز صفرم تا به روز دوم را جدا جدا محاسبه کنیم. بنابراین دو سناریو برای آرام بودم دومین روز وجود دارد:

ابتدا اینکه پس از روز صفرم, دو روز پیش رو هر دو آرام باشند که به احتمال زیر می تواند برقرار باشد.

(0.9)*(0.9) = 0.81

سپس در سناریو دیگر هوا می تواند روز بعد طوفانی و سپس دوباره آرام شود که احتمال بروز آن به شکل زیر است:

(0.1)*(0.5) = 0.05

بنابراین احتمال آرام بودن روز دوم به شرط آرام بودن اولین روز به شکل زیر می باشد:

P(S2 = C | S0 = C) = 0.81 + 0.05 = 0.86 or 86%

منبع: تغییر یافته تمرین کورس ai در washington

2. مساله تصویر بر روی ابرصفحه! نقطه دلخواه $c \in R^n$ و ابرصفحه به معادله $a^T.x + b = 0$ داده شده است که $a, x \in R^n$. هدف پیدا کردن نزدیک ترین نقطه این ابرصفحه نسبت به نقطه c است. این مساله را به صورت یک مساله محدب مقید مدل کرده و سپس حل کنید. (15 نمره)

$$min \quad \frac{1}{2}||x - c||^2_2$$
$$s.t \quad a^T x + b = 0$$

$$L(x, \lambda) = \frac{1}{2}||x - c||^2 + \lambda(a^T x + b)$$
$$\frac{dL}{dx} = (x - c) + \lambda a = 0 \rightarrow x = c - \lambda a$$
$$L(\lambda) = \frac{1}{2}\lambda^2.||a||^2 + \lambda(a^T.c - \lambda||a||^2 + b)$$
$$\frac{dL}{d\lambda} = \lambda||a||^2 + a^T.c - 2\lambda||a||^2 + b = 0 \rightarrow \lambda = \frac{a^T c + b}{||a||^2}$$
$$\rightarrow x = c - \frac{a^T c + b}{||a||^2} a =$$

3- الف) نگاشت انقباضی (Contraction Mapping) چیست؟ رابطه کلی آن به چه صورت است؟ (10 نمره)

نگاشت انقباضی تابعی است که فضا را به خودش نگاشت می‌کند و خاصیت انقباض دارد. به عبارت دیگر، برای هر دو نقطه در فضا، فاصله بین تصاویر آن دو نقطه تحت نگاشت، کمتر از فاصله خود آن دو نقطه است. (۵ نمره)

$$\left|\left|\tau\vec{V}_1 - \tau\vec{V}_2\right|\right|_\infty \leqslant \alpha\left|\left|\vec{V}_1 - \vec{V}_2\right|\right|_\infty$$ (۵ نمره)

ب) ثابت کنید الگوریتم Value Iteration یک Contraction Mapping است. متغیر γ چه نقشی دارد؟ (۱۵ نمره)

ابتدائا دو تابع V1 و V2 را به شکل زیر فرض می کنیم:

$$\tau\vec{V}_1(s) = Max_a \sum_{s'} p(s' \mid s, a)\left[R(s) + \gamma V_1(s')\right]$$
$$\tau\vec{V}_2(s) = Max_a \sum_{s'} p(s' \mid s, a)\left[R(s) + \gamma V_2(s')\right]$$

در ادامه اکشن بهینه V1 را در هر دو قرار داده و دو عبارت را از هم کم می‌کنیم:

$$\tau\vec{V}_1(s) - \tau\vec{V}_2(s) \le \sum_{s'} P(s'|s,a)\left[R(s) + \gamma V_1(s')\right] - \sum_{s'} P(s'|s,a)\left[R(s) + \gamma V_2(s')\right]$$
$$= \gamma \sum_{s'} P(s'|s,a)\left[V_1(s') - V_2(s')\right] \le \gamma\, Max_{s'}\left|\left|\vec{V}_1(s') - \vec{V}_2(s')\right|\right| \sum_{s'} P(s'|s,a)$$
$$= \gamma\, Max_{s'}\left|\left|\vec{V}_1(s') - \vec{V}_2(s')\right|\right|$$
$$\Rightarrow \tau\vec{V}_1(s) - \tau\vec{V}_2(s) \le \gamma\, Max_{s'}\left|\left|\vec{V}_1(s') - \vec{V}_2(s')\right|\right|$$

با مقایسه فرم نهایی به دست آمده با Contraction Mapping متغیر γ معادل α است.

ج) ثابت کنید الگوریتم Value Iteration به عنوان یک Contraction Mapping همواره همگرا می‌شود. (۱۵ نمره)

$$\left|\left|\vec{V}_k - \vec{V}^*\right|\right|_\infty = \left|\left|\tau\vec{V}_{k-1} - \tau\vec{V}^*\right|\right|_\infty \le \gamma\left|\left|\vec{V}_{k-1} - \vec{V}^*\right|\right|_\infty$$
$$\le \gamma^2\left|\left|\vec{V}_{k-2} - \vec{V}^*\right|\right|_\infty \le \ldots \le \gamma^k\left|\left|\vec{V}_0 - \vec{V}^*\right|\right|_\infty = \gamma^k A$$
$$0 \le \left|\left|\vec{V}_k - \vec{V}^*\right|\right|_\infty \le \gamma^k A, \ \lim_{k\to\infty} \gamma^k A = 0$$
$$\Rightarrow \left|\left|\vec{V}_k - \vec{V}^*\right|\right|_\infty \to 0$$

منبع: اثبات کلاسی

4. دو استراتژی متفاوت یادگیری در یادگیری تقویتی روش های Temporal Difference و Monte Carlo هستند.

الف) این دو را در نحوه آپدیت policy با یکدیگر مقایسه کنید. (۱۰ نمره)

روش Monte Carlo صبر میکند تا یک episode پایان یابد تا دنباله state , action ,reward های آن را دریافت کند. سپس مقادیر value function هر state را بر اساس ریوارد واقعی کسب شده از محیط آپدیت میکند. (۵ نمره)

در مقابل Temporal Difference به ازای هر تغییر وضعیت، value function استیت مبدا را آپدیت میکند. این آپدیت تابعی از ریوارد آنی دریافت شده از محیط و estimated value استیت مقصد است. (5 نمره)

ب) یکی از مسائل مهم در انواع الگوریتم های یادگیری ماشین trade-off میان Bias و Variance است. نحوه ظهور این ۲ مسئله را در value estimate بیان کنید. (۱۰ نمره)

در یادگیری تقویتی، یک تخمین دارای variance بالا، value estimate های پراکنده اما با میانگین نزدیک به درست را نتیجه می دهد. در حالی که تخمین های دارای Bias بالا, value estimate های نسبتا stable اما همواره دارای خطا را تولید می کند. برای مشخص تر کردن این موضوع، یک بازی دارت را تصور کنید. یک بازیکن با Bias بالا کسی است که همیشه در نزدیکی هدف ضربه می زند، اما به طور پیوسته در یک جهت فاصله دارد. از طرف دیگر یک بازیکن با variance بالا، کسی است که گاهی اوقات به هدف ضربه می زند، و گاهی اوقات خارج می شود، اما به طور متوسط در نزدیکی هدف است (۵ نمره برای توضیح هر مورد)

ج) این تقابل را در روش های Temporal Difference و Monte Carlo مقایسه و بررسی کنید. (۱۵ نمره)

روش Monte Carlo مقدار value function هر استیت را میانگین مقادیر اپیزود های مختلفی که شامل آن می‌شده در نظر میگیرد. در این حالت, همانطور که یک استیت میتواند در مسیر رسیدن به گل قرار داشته باشد همچنین می تواند در مسیر یک هدف مرگبار نیز باشد. از طرفی Stochastic بودن محیط نیز می تواند باعث تنوع در ریوارد دریافتی از یک مسیر واحد شود. (5 نمره) این موضوع سبب می شود تا مقادیر دارای variance بالا به یک استیت نسبت داده شود که Monte Carlo را یک روش high-variance میکند. (۲.۵ نمره)

از سوی دیگر Temporal Difference مقدار هر استیت را بر اساس ریوارد حاصل از Action انجام شده بر روی آن و value تخمین زده استیت بعدی آپدیت می کند که سبب می شود پدیده نامطلوب قبلی رخ ندهد. اما از سوی دیگر در ابتدای فرآیند مجبور به مقدار دهی اولیه (هرچند صفر) به هر state هستیم که از درستی آن اطلاعی نداریم. این عمل سبب تولید Bias در تخمین ها می شود اما در بسیاری موارد میتوان با داشتن تعداد زیادی اپیزود مختلف آن را در نظر نگرفت. مشکل جدی تر جایی رخ میدهد که برای هر آپدیت مجبور به تخمین value استیت بعدی هستیم(۵ نمره) و این عامل سبب ایجاد Bias بالا در تخمین نهایی می شود.(۲.۵ نمره)
منبع۱, منبع۲, منبع۳, منبع۴