هوالحکیم

دانشگاه تهران
پردیس فارابی
دانشکده‌ی مهندسی

هوش مصنوعی پیشرفته
نیمسال دوم ۹۶-۱۳۹۵
http://courses.fouladi.ir/aai

تکلیف شماره‌ی ٤

فصل بیستم و بیست‌ویکم

یادگیری: یادگیری مدل‌های احتمالاتی / یادگیری تقویتی

Learning: Learning Probabilistic Models / Reinforcement Learning

۱) دو آماردان به پزشک می‌روند و به هر دوی آنها از قبل در مورد پیشرفت بیماری اطلاع داده شده است: به احتمال ۴۰ درصد بیماری آنها بیماری A است که کشنده است و به احتمال ۶۰ درصد بیماری B است که وخیم است. خوشبختانه، داروهای ارزان قیمت ضد بیماری A و ضد بیماری B وجود دارند که ۱۰۰ درصد مؤثر هستند و هیچ عارضه‌ی جانبی هم ندارند. آماردانان حق انتخاب در مصرف یک دارو یا هر دو و یا هیچ‌کدام را دارند. اولین آماردان که به نظریه‌ی بیزی گرایش دارد، چه خواهد کرد؟ آماردان دوم که همیشه از فرضیه‌ی دارای ماکزیمم درست‌نمایی استفاده می‌کند، چه خواهد کرد؟
پزشک با انجام تحقیقات کشف می‌کند که در واقع بیماری B دو نوع دارد: دیکسترو B و لِوو B که احتمال درمان آنها توسط داروی ضد بیماری B به یک اندازه است. اکنون که سه فرضیه وجود دارد، دو آماردان چه کاری انجام خواهند داد؟

۲) m نقطه داده‌ی آموزشی (x_j, y_j) را در نظر بگیرید که در آن y_jها از روی x_jها بر اساس مدل خطی گاوسی مطابق معادله‌ی زیر تولید می‌شوند:

$$P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$$

مقادیر θ_1، θ_2، و σ که لگاریتم درست‌نمایی شرطی را ماکزیمم می‌کند، بیابید.

۳) این تمرین، خصوصیات توزیع بتا (که در یادگیری پارامتر بیزی کاربرد دارد) را بررسی می‌کند.

(الف) با انتگرال‌گیری روی بازه‌ی $[0, 1]$ نشان دهید که ضریب نرمال‌سازی برای توزیع $\text{beta}[a, b]$

$$\text{beta}[a, b](\theta) = \alpha\theta^{a-1}(1-\theta)^{b-1}$$

با رابطه‌ی

$$\alpha = \Gamma(a+b)/\Gamma(a)\Gamma(b)$$

داده می‌شود که در آن $\Gamma(x)$ تابع گاما است که با رابطه‌ی $\Gamma(x+1) = x \cdot \Gamma(x)$ و $\Gamma(1) = 1$ تعریف می‌شود. (برای عدد صحیح x داریم: $\Gamma(x+1) = x!$.)

(ب) نشان دهید که میانگین این توزیع، $a/(a+b)$ است.

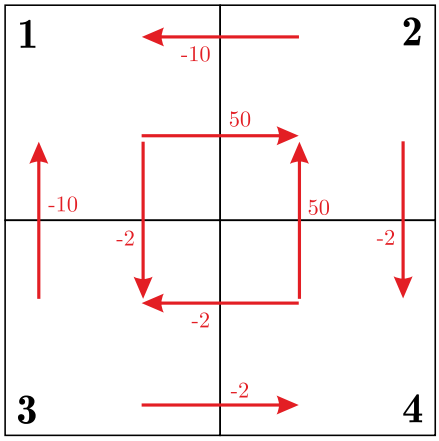
(ج) مد(های) این توزیع (محتمل‌ترین مقدار(های) θ) را بیابید.

۴) مزیت یادگیری تفاضل زمانی ارزش‌های Q (Q-learning) بر یادگیری تفاضل زمانی ارزش حالت‌ها V چیست؟

۵) در برخی مسائل یادگیری تقویتی، پاداش‌ها (rewards) برای حالت‌های هدف مثبت، و برای سایر موارد صفر یا منفی است. آیا علامت پاداش‌ها مهم است؟ یا تنها فاصله‌ی میان آنها اهمیت دارد؟ با استفاده از تعریف پاداش تخفیف‌یافته R_t (discounted reward) ثابت کنید که افزودن یک مقدار ثابت به همه‌ی پاداش‌های ابتدایی، باعث می‌شود ثابت K به ارزش همه‌ی حالت‌ها افزوده شود و بنابراین، بر ارزش‌های نسبی همه‌ی حالت‌ها تحت همه‌ی سیاست‌ها تأثیری نمی‌گذارد. مقدار K را بر حسب C و نرخ تخفیف (γ) محاسبه کنید.

۶) همگرایی الگوریتم Q-learning را ثابت کنید. برای این منظور، فرض کنید حالت دنیا قطعی (deterministic) است و هر جفت حالت-کنش (s, a) به تعداد دفعات نامتناهی مشاهده می‌شود. یک بازه ی کامل (interval) را به‌عنوان بازه‌ای در نظر بگیرید که در خلال آن همه‌ی جفت‌های ممکن (s, a) مشاهده شده است. نشان دهید که در خلال هر چنین بازه‌ای، مقدار قدر مطلق ماکزیمم خطا در جدول Q با ضریب γ کاهش می‌یابد. در نتیجه، برای $\gamma < 1$ پس از بی‌نهایت به‌هنگام‌سازی، ماکزیمم خطا به سمت صفر میل می‌کند.

۷) شکل زیر، یک دنیای شبکه‌ای چهار خانه‌ای (هر خانه معادل با یک حالت) را به تصویر کشیده است که در آن حالت 2 «طلا» را نشان می‌دهد. با استفاده از مقادیر پاداش بلافصل (immediate reward) نشان داده‌شده بر روی پیکان‌ها ($R(s,a,s')$) و با به‌کارگیری الگوریتم Q-learning بر روی حالت‌ها به‌صورت پادساعتگرد حرکت کنید و جدول state-action را تا ۴ دور به‌روزرسانی کنید. فرض کنید نرخ تخفیف $\gamma = ۰/۹$ باشد.



۸) یک ربات متحرک خودمختار را در نظر بگیرید که می‌تواند در هرگام زمانی، سریع (FAST) یا کند (SLOW) حرکت کند. حرکت FAST به طورکلی دارای پاداش $+۲$ است در حالی که حرکت SLOW تنها پاداش $+۱$ را دارد. در هر صورت، ربات باید دمای داخلی خودش را در نظر بگیرد که HOT یا OK می‌باشد. حرکت کردن در حالت SLOW منجر به دمای کمتری می‌شود، در حالی که حرکت کردن در حالت FAST منجر به افزایش دما می‌شود. اگر ربات HOT باشد، این خطر وجود دارد که overheat شود. در این نقطه باید بایستد، دمایش را کم کند و تعمیر شود. گذارهای MDP و پاداش‌های متناظر مطابق جدول زیر مشخص می‌شود:

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
OK	SLOW	OK	۱/۰	$+۱$
OK	FAST	OK	۰/۵	$+۲$
OK	FAST	HOT	۰/۵	$+۲$
HOT	SLOW	OK	۱/۰	$+۱$
HOT	FAST	HOT	۰/۵	$+۲$
HOT	FAST	OK	۰/۵	-۱۰

توجه کنید که با پرداخت هزینه‌ی تعمیر (سطر آخر)، ربات پس از آن OK می‌شود.

(الف) دو تکرار از الگوریتم «تکرار ارزش» را با در نظر گرفتن نرخ تخفیف $\gamma = ۰/۸$ در جدول زیر اجرا کنید و فقط مقادیر دایره را مشخص کنید.

s	$V_۰$	$V_۱$	$V_۲$
OK	0	◯	◯
HOT	0	◯	

(ب) الگوریتم Q-learning را با نرخ تخفیف $\gamma = ۰/۸$ اجرا کنید. نرخ یادگیری را $\alpha = ۰/۵$ در نظر بگیرید و از نمونه گذارهای زیر استفاده کنید. مقادیری از Q که در طول یک گام تغییر نمی‌کنند را بازنویسی نکنید. مقادیر اولیه صفر هستند. نمونه‌ی گذارها در تجربه‌ی عامل عبارتند از:

(OK, FAST, HOT), reward $= +۲$, (calculate $Q_۱$)
(HOT, FAST, OK), reward $= -۱۰$, (calculate $Q_۲$)
(OK, SLOW, OK), reward $= +۱$, (calculate $Q_۳$)