دانشگاه تهران
پردیس فارابی
دانشکده‌ی مهندسی

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

تکلیف شماره‌ی ۳

فصل شانزدهم و هفدهم

تصمیم‌گیری: اتخاذ تصمیم‌های ساده، اتخاذ تصمیم‌های پیچیده

DECISION MAKING: MAKING SIMPLE DECISIONS / MAKING COMPLEX DECISIONS

۱) در ۱۸۳۷م، نیکولاس برنولی، معمایی را با نام پارادوکس سنت پترزبورگ بیان کرد که به صورت زیر بیان می‌شود: «سکه‌ی سالمی انداخته می‌شود و شما تا زمانی که شیر بیاید فرصت دارید بازی کنید. اگر اولین شیر آمدن در n-امین پرتاب اتفاق بیفتد، ۲^n دلار جایزه می‌گیرید.»

(الف) نشان دهید که امید ارزش پولی (EMV) این بازی، بی‌نهایت است.

(ب) شما چه‌قدر حاضرید برای ورود به این بازی پرداخت کنید؟

(ج) برنولی این پارادوکس را با این پیشنهاد حل کرد که سودمندی پول با مقیاس لگاریتمی اندازه‌گیری شود، یعنی $U(S_n) = a\log_۲ n + b$ که در آن S_n حالت داشتن n دلار است. متوسط سودمندی این بازی تحت این فرض چه‌قدر است؟

(د) با فرض سرمایه‌ی اولیه‌ی k دلار، حداکثر مبلغ پرداختی برای ورود به این بازی که رسیونال باشد، چه‌قدر است؟

۲) فرض کنید سودمندی یک دنباله از حالت‌ها، ماکزیمم پاداش به‌دست آمده در میان تمام حالت‌های آن دنباله تعریف شود. نشان دهید که این تابع سودمندی موجب ترجیحات ایستان(stationary preferences) میان دنباله‌های حالت‌ها نمی‌شود. آیا با این وجود ممکن است که یک تابع سودمندی بر روی حالت‌ها تعریف کنیم که اتخاذ تصمیم MEU موجب رفتار بهینه شود؟

۳) آیا هر مسئله‌ی جستجوی متناهی را می‌توان دقیقاً به یک مسئله‌ی تصمیم‌گیری مارکوف تبدیل کرد به‌طوری که یک راه‌حل بهینه‌ی آن، یک راه‌حل بهینه‌ی مسئله‌ی اول باشد؟ اگر چنین است، با دقت چگونگی تبدیل مسئله و بازگرداندن راه‌حل مسئله‌ی تبدیل شده را توضیح دهید. در غیر این صورت، با دقت توضیح دهید که چرا نه (مثلاً یک مثال نقض بزنید).

۴) فرض کنید یک MDP بدون تخفیف (undiscounted) دارای ۳ حالت (۱, ۲, ۳) به‌ترتیب با پاداش‌های (−۱, −۲, ۰) باشد. حالت ۳ یک حالت پایانی است. در حالت‌های ۱ و ۲ دو کنش ممکن وجود دارد: $\{a, b\}$. مدل گذار به‌صورت زیر است:

- در حالت ۱ کنش a به‌احتمال ۰٫۸ عامل را به حالت ۲ منتقل می‌کند و با احتمال ۰٫۲ عامل را همان‌جا باقی می‌گذارد.
- در حالت ۲ کنش a به‌احتمال ۰٫۸ عامل را به حالت ۱ منتقل می‌کند و با احتمال ۰٫۲ عامل را همان‌جا باقی می‌گذارد.
- در حالت‌های ۱ و ۲ کنش b به‌احتمال ۰٫۱ عامل را به حالت ۳ منتقل می‌کند و با احتمال ۰٫۹ عامل را همان‌جا باقی می‌گذارد.

حال به پرسش‌های زیر پاسخ بدهید:

(الف) به‌صورت کیفی، چه چیزی در مورد سیاست بهینه در حالت‌های ۱ و ۲ می‌تواند تعیین شود؟

(ب) از الگوریتم تکرار سیاست استفاده کنید و سیاست بهینه را به‌صورت کمی ارزیابی کنید تا ارزش‌های حالت‌های ۱ و ۲ مشخص شود. فرض کنید سیاست آغازین در هر دو حالت کنش b را در نظر گرفته است.

(ج) اگر سیاست آغازین در هر دو حالت کنش a را در نظر گرفته باشد، با تکرار سیاست چه اتفاقی خواهد افتاد؟ آیا تخفیف (discounting) کمکی می‌کند؟ آیا سیاست بهینه به فاکتور تخفیف بستگی دارد؟

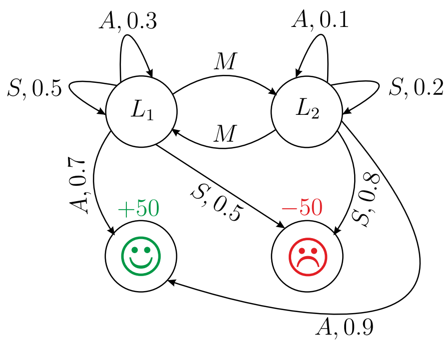
۵) گاهی اوقات، MDPها با یک تابع پاداش $R(s, a)$ فرمول‌بندی می‌شوند که به کنش انجام شده a در حالت s وابسته است یا با یک تابع پاداش $R(s, a, s')$ که به حالت برآمد s' هم وابسته است.

(الف) معادلات بلمن را برای این فرمول‌بندی‌ها بنویسید.

(ب) نشان دهید که یک MDP با تابع پاداش $R(s,a,s')$ می‌تواند به یک MDP دیگر با تابع پاداش $R(s,a)$ تبدیل شود، به‌گونه‌ای که سیاست‌های بهینه در MDP جدید دقیقاً متناظر با سیاست‌های بهینه در MDP اصلی باشد.

(ج) کار مشابهی را برای تبدیل MDPهایی با $R(s,a)$ به MDPهایی با $R(s)$ انجام بدهید.

۶) نمودار زیر یک مدل MDP از یک دعوای خیابانی (!) را ترسیم کرده است.



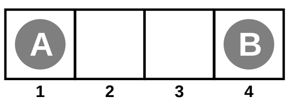
در این دعوا می‌توان بین محل‌های $L_۱$ و $L_۲$ جابه‌جا شد. یکی از این مکان‌ها به حریف نزدیک‌تر است. اگر عامل از نزدیک‌ترین حالت حمله کند، شانس بیشتری برای موفقیت دارد (۰٫۹) (در مقابل ۰٫۷ برای مکان دورتر از حریف). به‌هرحال، ممکن است عامل توسط حریف دیده شود و مضروب شود (با شانس ۰٫۸) (در مقابل شانس ۰٫۵ برای مکان دورتر). اگر عامل در یک مکان بماند (stay) ممکن است توسط حریف دیده شود. عامل نیازمند یک سیاست بهینه برای کنش در این محیط نامطمئن است. در گراف فوق، پیکان‌ها کنش‌های ممکن را نشان می‌دهد (M کنش قطعی حرکت move، A کنش تصادفی حمله attack، S کنش تصادفی ماندن stay). بر روی یال‌ها (a, p) نوشته شده است (کنش و احتمال گذار حالت). همه‌ی پاداش‌ها در همه‌ی مراحل صفر هستند، به‌جز حالت‌های پایانی که موفقیت عامل در آن با پاداش ۵۰+ و موفقیت حریف با پاداش ۵۰− برای عامل ما نشان داده شده است. به استفاده از فاکتور تخفیف $\gamma = ۰٫۹$ سیاست بهینه را محاسبه کنید.

۷) در این تمرین، MDPهای دو بازیکنه را بررسی می‌کنیم. فرض می‌کنیم بازیکنان A و B باشند. $R(s)$ پاداش برای بازیکن A در s است و پاداش B همیشه قرینه‌ی پاداش بازیکن A است (بازی مجموع صفر).

(الف) فرض می‌کنیم $U_A(s)$ سودمندی حالت s زمانی که نوبت A برای حرکت در حالت s است و $U_B(s)$ سودمندی حالت s زمانی که نوبت B برای حرکت در حالت s است .تمام پاداش‌ها و سودمندی‌ها از دیدگاه A محاسبه می‌شوند (درست مانند درخت بازی MiniMax). معادلات بلمن را با تعریف $U_A(s)$ و $U_B(s)$ بنویسید.

(ب) چگونگی انجام تکرار ارزش دو بازیکنه را با این معادلات توضیح دهید و یک معیار مناسب برای توقف بازی ارائه دهید.

(ج) یک بازی دونفره را در نظر بگیرید که بر روی یک صفحه با چهار موقعیت که از ۱ تا ۴ شماره‌گذاری شده و در طول یک خط قرار گرفته‌اند انجام می‌شود. هر بازیکن یک نشانه دارد. بازیکن A با نشانه‌ی خود از موقعیت ۱ و بازیکن B با نشانه‌ی خود از موقعیت ۴ آغاز می‌کند. ابتدا بازیکن A حرکت می‌کند.



این دو بازیکن به نوبت حرکت می‌کنند. هر بازیکن باید نشانه‌ی خود را از یک خانه به فضای خالی موجود در یکی از دو جهت انتقال دهد. اگر رقیب یکی از خانه‌های مجاور را اشغال کرده باشد، بازیکن می‌تواند از روی رقیب پرش کند و در صورت وجود خانه‌ی خالی در محل بعدی قرار گیرد (برای مثال اگر A در ۳ باشد و B در ۲ باشد، آنگاه A می‌تواند به ۱ برگردد). بازی وقتی تمام می‌شود که یکی از بازیکن‌ها به موقعیت مقابل خود بر روی صفحه برسد. اگر بازیکن A ابتدا به فضای ۴ برسد، در این صورت ارزش بازی ۱+ و اگر بازیکن B ابتدا به فضای ۱ برسد، در این صورت ارزش بازی ۱− خواهد بود. فضای حالتی (به‌جای درخت بازی) را ترسیم کنید که حرکت‌های A را با خطوط ممتد و حرکت‌های B را با خطوط مقطع نشان دهد. هر حالت را با $R(s)$ علامت‌گذاری کنید. مرتب کردن حالت‌های (s_A, s_B) در یک شبکه‌ی دوبعدی با استفاده از مختصات (s_A, s_B) مفید خواهد بود.

(د) الگوریتم تکرار ارزش دو بازیکنه را برای حل این بازی به‌کار بگیرید و سیاست بهینه را استخراج کنید.