«فصل یک : سیستم های یادگیری»

ما بطور کلی ۵ روش یادگیری در ML داریم : ① Supervised learning : نتایج مطلوب به آن داده می‌شود، متناسب با آن آموزش

صورت می‌گیرد.

② unSupervised learning : نتایج مطلوب در اختیار agent قرار نمی‌گیرد، agent با کشف ویژگی های مشترک بین آنها

طبقه بندی می کند.

③ Reinforcement learning (RL) : agent در محیط به دنبال برآورده کردن هدفی است که در مراحل (step) زمانی متناوب

با state محیط، یک action (عمل) را انتخاب می کند و به محیط اعمال می کند و سپس براساس پاداش یا مجازات دریافتی از محیط وظیفه خود

را یاد می گیرد.

④ Semi_Supervised learning : بدون نظارت است که تعداد محدودی از پاسخ ها نیز توسط یک معلم در اختیار agent قرار می گیرد

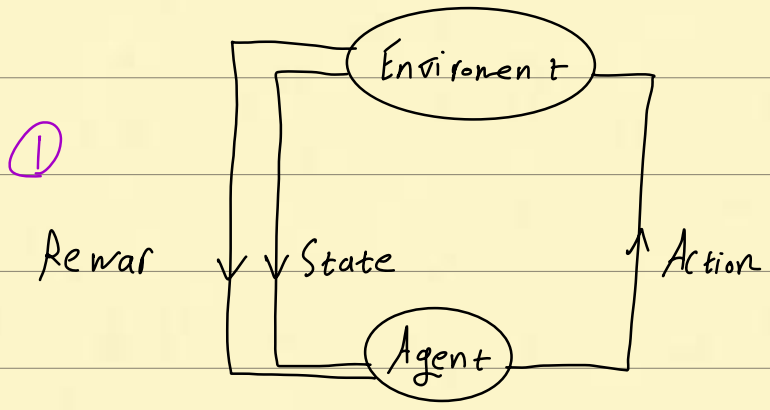
و agent از این پاسخ های محدود برای بهبود پروسه یادگیری بدون نظارت استفاده می کند

⑤ Self_supervised learning : معلم وجود ندارد و خود agent خود به خود با توجه به تکنیک های خود را آموزش می دهد.

* AGI (Artificial general intellegent) : هدف آن ترکیب روش های بالا و ربات انسان نما است.

اما بحث ما اینجا RL است که در صفحه بعد به صورت کلی نمایی از آن آمده است :

نمای کلی یک مسئله RL بصورت مقابل:

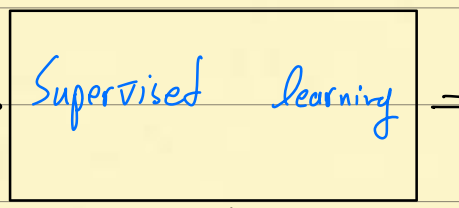


① تفاوت RL با یادگیری با نظارت ←

در یادگیری با نظارت براساس ورودی خروجی آموزش داده می‌شود که در RL صرفاً براساس پاداش یا تنبیه است و agent درجهت ماکسیمم کردن پاداشی در زمان آینده تصمیم را انجام می‌دهد.

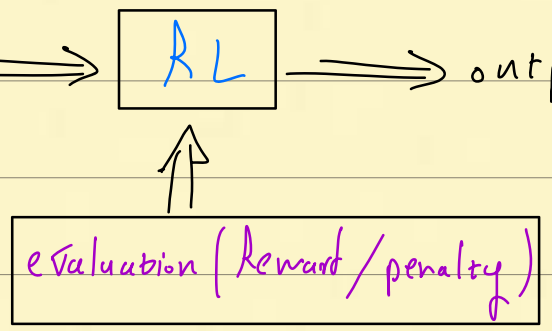
② در یادگیری با نظارت بصورت Offline است ولی در RL یادگیری بصورت Online است.

a) Supervised learning:



هدف؟ خنثی کردن خطای بین Real و desired است.

b) Reinforcement learning:

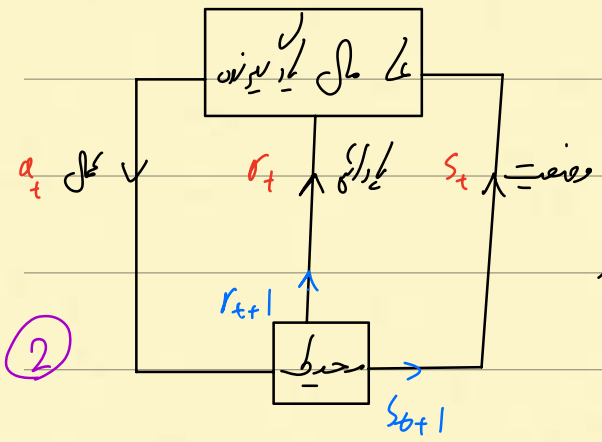


هدف؟ جمع کردن حداکثر پاداش ممکن در زمان معین است.

* در بعضی پارادایم های آینده و انتهای بازه زمانی صحیح تر است .

فصل دوم ⇐ بخش های اساسی یک سیستم RL

در شکل کلی یک رویکرد RL عامل یادگیرنده (agent) با توجه به عامل S_t و پارادایم r_t



در بازه زمانی قبل ورودی عمل a_t را تولید و به محیط اعمال کرده و محیط در پاسخ state های جدید

و پارادایم های جدید به ترتیب S_{t+1} و r_{t+1} تولید می کند.

بخش های سیستم RL به طور کلی در کل ۵ یا ۶ بخش است :

① محیط (environment) ② عامل یادگیرنده (Agent) ③ سیاست (policy)

④ تابع پارادایم (Reward function) ⑤ تابع ارزیابی (Value function)

* در برخی سیستم های RL یک مدل از محیط نیز داریم.

policy (سیاست) ⇐ نحوه رفتار کردن عامل یادگیرنده را مشخص می کند ، نگاشتی از حالتهای محیط به عملهایی است که در آن حالتها می توانند

انتخاب بشوند و ۳ حالت کلی دارد:

① یک تابع دقیق و کافی است . ② یک جدول جست و جو است .

③ تابع احتمالاتی از اعمال

«Reward (پاداش)» ← این تابع هر حالت یا state از محیط را به یک عدد اسکالر، یک پاداش نسبت می‌کند.

این عدد میزان رضایت محیط از وارد شدن به آن state را نشان می‌دهد.

«Value (ارزش)» ← برخلاف Reward و یا پاداش تعیین می‌کند در بلند مدت چه چیزی سودمند است.

به عبارتی دیگر ارزش هر state کل مقدار پاداشی را معین می‌کند که یک عامل می‌تواند انتظار داشته باشد با شروع از آن در آینده جمع آوری کند.

* امکان دارد عملی پاداش لحظه‌ای (r) کمی داشته باشد اما ارزش آن بالا باشد چرا که در آینده برای agent پاداش‌های بیشتری دارد.

* Value یا ارزش وابسته به سیاست است که از آینده داریم و تحت یک سیاست بیان می‌کنیم.

* به دست آوردن Value (ارزش) بسیار سخت‌تر از محاسبه Reward (پاداش) است چون پاداش مستقیم از محیط دریافت می‌شود ولی ارزش را باید تخمین زد.

«روش‌های حل مسائل RL» ← ① جست‌وجوی سیاست: تابع سیاست‌های ممکن جست‌وجو می‌شود و الگوریتم‌های ML بدون آنکه از این تابع ارزش استفاده کنیم و یا درک درستی از محیط داشته باشیم بهترین سیاست را می‌یابد.

* هرچه فضا بزرگ‌تر حجم محاسبات سنگین‌تر

مثال) جست‌وجوی تمام حرکات ممکن در یک بازی شطرنج.

② تکامل با محیط : سعی در پیدا کردن سیاست بهینه با تعامل با محیط است.

روشهای دیگر قبیل از: • برنامه ریزی دینامیک ← مبتنی بر مدل، دقیق ولی محاسبات بالا هستند.

• مونت کارلو ← مبتنی بر تجربه agent است (غیر وابسته به مدل)، دقیق نیستند ولی ارزیاب هستند.

• تفاضلی ← ترکیب ۲ مدل بالا است که خود شامل Q-learning و SARSA و ... است.

«مدل محیط» ← یک مدل محیط خوب می‌تواند رفتار محیط را پیش بینی کند که در برنامه ریزی دینامیک بحث می‌شود.

«فرم مسئله Tic_Tac_Toe» ← action : انتخاب هرخانه ، state ← کجا و وضعیت X , O

Reward : کجا سطر و ستون و ردیف است

«فصل سوم ← انواع ساختار های هوشمند RL»

agent ← هرچیزی که قادر است محیط پیرامون خود را ازطریق حسگرهایی درک کند و بر محیط توسط عملگرهایی عمل کند.

ما بصورت کلی ۲ نوع agent داریم:

① agent هایی که از قبل مدل محیط به آنها داده نشده است اما تابع ارزیابی (value func) را دارند.

* ۲ نوع نشان دادن معروف داریم: • $R^a_{ss'}$ یا $r(s_t, a_t)$ ← مقدار پاداشی دریافت شده از اعمال action به محیط در حالت s

ورفتن به حالت s' است.

- $P_{ss'}^{a}$ یا $Pr(S_{t+1}|S_t, a_t)$ ⇐ احتمال رفتن محیط به حالت s' در صورت عمل a_t از حالت s است.

حال در مدل اول agent ما $P_{ss'}^{a}$ و $R_{ss'}^{a}$ را داریم.

این مدلها برای مدلی که از محیط دارند تابع ارزشی را به طور مستقیم تخمین می‌زنند و از آنجایکه مدل محیط همهٔ مشخصات مورد نیاز از محیط

را در بر دارد، لذا برای تخمین تابع ارزشی (Value func) نیاز تعامل مستقیم با محیط نداریم.

پس تخمین تابع ارزشی بصورت offline (جدای از محیط) است.

(2) agent هایی که هیچکدام از مولفه‌های محیط یا تابع ارزشی به آنها داده نشده است. پس بنابراین نیاز دارند

تا بصورت مستقیم با محیط تعامل کنند و بتوانند تابع ارزشی را تخمین بزنند.

این کار یعنی تخمین تابع ارزشی به ۲ صورت است:

- مبتنی بر مدل ⇐ اول agent با تعامل همراه محیط یک مدل از محیط را آموزش می‌بیند و سپس به کمک آن مدل

تابع ارزشی را بصورت آنلاین تخمین می‌زند که روشی Adaptive programming یا ADP است.

- مستقل از مدل ⇐ بدون اینکه مدلی از محیط بدست آید agent بدون واسطه و از طریق تعامل مستقیم با محیط

به یادگیری تابع ارزشی می‌پردازد که روشهای Mont Carlo یا Tempored difference هستند.

«انواع محیط های تسک های RL»

1) «محیط مارکوف (Markof)» ← محیطی که در آن پیش بینی وضعیت آینده صرفا با آخرین وضعیت (State) ما وابسته است، مثال: بازی شطرنج، کنترل ماشین، کنترل هواپیما

* بعضی از کار ها در محیط غیر مارکوف است مثال پیش بینی رفتار انسان، پیش بینی آب و هوا و...

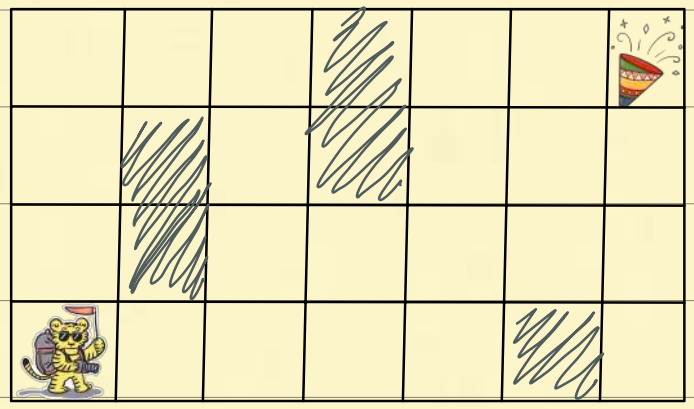
2) «محیط تسک / پیوسته» ← محیطی که در آن فضای حالت - عمل (action - State) تسک یا پیوسته می باشد.

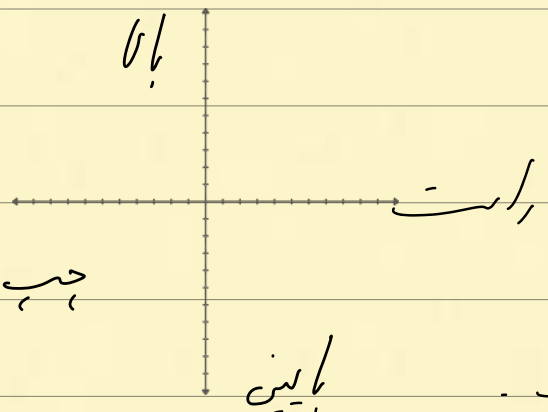
مثال: در هزارتو یا راهروی ماریچ (Maze) محیطی با ابعاد 4x7 داریم که حاوی یک تنبع در یکی از سلول‌ها و یک agent جست‌وجو کننده در یکی از خانه‌ها دارد همچنین تعدادی مانع نیز در مسیر قرار داده شده است.

اعمالی که باعث (خروج agent از مسیر می‌شود یا باعث برخورد به مانع ها می‌شود را پاداشی ۱- = r) و (رسیدن به تنبع پاداشی ۱+ = r) و سایر پاداش ها ۰ = r) است.

همانطور که در شکل مشخص است این اعمال تسک هستند و به بالا، پایین، چپ و راست خلاصه می‌شوند.

3)





* این یک محیط مارکوف است.