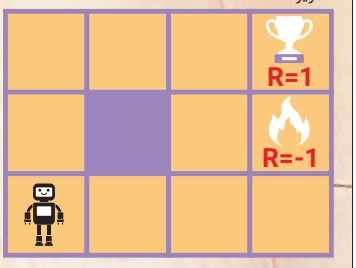
آشنایی با یادگیری تقویتی

بیشترین پاداش را از محیط بگیر

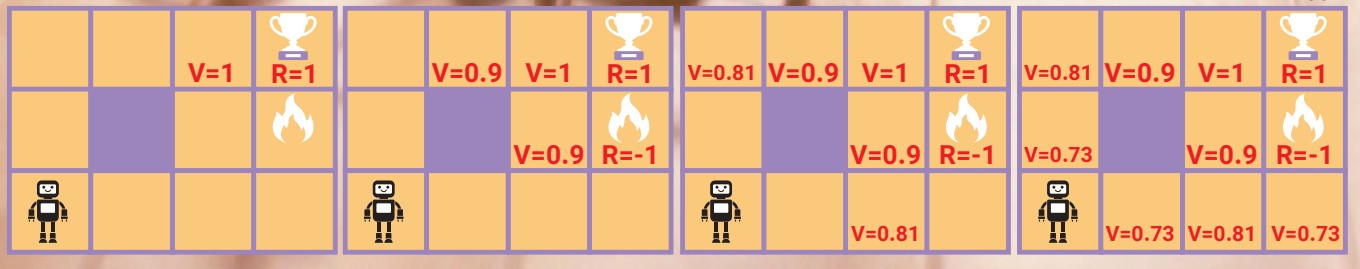
نوع دیگری از یادگیری ماشین، یادگیری نیمه نظارتی یا یادگیری تقویتی است که از فرایند یادگیری انسان الهام گرفته شده است. هنگامی‌که کودک تلاش می‌کند برای اولین بار دوچرخه‌سواری بیاموزد، در تجربه‌های اول که مدام زمین می‌خورد، با شکست‌های متوالی مواجه خواهد شد. در هر شکست، مغز کودک یاد می‌گیرد خود را با محیط و شرایط محیطی تطبیق دهد، به‌گونه‌ای که بتواند تعادل کودک را حفظ کند. با تمرین و تکرار عضلات شکل هماهنگ و لازم برای حفظ تعادل را یاد می‌گیرند. هنگامی‌که فرایند یادگیری به اتمام می‌رسد، دوچرخه‌سواری مانند راه‌رفتن یا نفس‌کشیدن به‌صورت کاملاً خودکار انجام می‌شود.

در این مثال، کودک به‌عنوان یک عامل هوشمند شناخته می‌شود که با تعامل با محیط قرار است یاد بگیرید دوچرخه‌سواری را به‌درستی انجام دهد. محیط مانند یک معلم است که نتیجهٔ درستی یا نادرستی انجام عمل را از طریق پاداش یا جریمه به عامل بر می‌گرداند. فرض کنید مجموعه اعمالی که کودک می‌تواند انجام دهد، شامل «رکاب زدن، ترمزکردن، مایل‌شدن به چپ و مایل‌شدن به راست» باشد! کودک در ابتدا در حالت توقف قرار دارد و در هر لحظه باید یکی از اعمال را انتخاب کند. پس از انتخاب هر عمل، کودک از یک حالت به حالت دیگر می‌رود. فرض کنید کودک در حالت توقف است و عمل «مایل‌شدن به راست» را انتخاب می‌کند. در نتیجهٔ این عمل، کودک به زمین می‌خورد. زمین‌خوردن کودک جریمه‌ای است که محیط برای انجام یک عمل نامناسب (مایل‌شدن به راست) در آن حالت (حالت توقف) برای کودک در نظر می‌گیرد. در تلاش بعدی کودک یاد می‌گیرد در حالت توقف نباید عمل «مایل‌شدن به راست» را انجام دهد، زیرا می‌داند که با جریمهٔ زمین‌خوردن مواجه می‌شود. فرایند یادگیری آن‌قدر ادامه می‌یابد که کودک یاد بگیرید در حالت توقف بهترین عملی که باید انجام دهد «رکاب‌زدن» است. هنگام رکاب‌زدن، دوچرخه به سمت جلو حرکت می‌کند، اما ممکن است قدری به سمت چپ یا راست متمایل شود. فرض کنید کودک در حالت متمایل به چپ قرار دارد. در این لحظه باید یاد بگیرید کمی به سمت راست متمایل شود تا در حالت تعادل باقی بماند. در حالت تعادل، محیط یک پاداش به او می‌دهد و این پاداش لذت دوچرخه‌سواری است. کودک

تصویر ۱



تصویر ۲



در هر حالتی که باشد، سعی می‌کند عملی را انتخاب کند که بیشـــترین پاداش را از محیط دریافت کند. یادگیری تقویتی به انتخاب بهترین عمل در هر حالت گفته می‌شود و هدف از یادگیری مشخص‌کردن حالت بهینه است.

بنابراین در یادگیری تقویتی، عامل می‌آموزد چگونه با تعامل با محیط، رفتار خود را بهبود دهد. این عامل یک موجود خودکار است که از طریق حسگرهایش محیط را درک می‌کند و می‌تواند با استفاده از محرک‌هایش اعمالی را انجام دهد. به‌طور مثال، یـــک ربات جمع‌آوری‌کنندهٔ زباله را در نظر بگیرید که با اســـتفاده از چرخ در محیط حرکت می‌کند. با استفاده از دوربین محیط اطراف خود را می‌بیند و از طریق بازوهای رباتی، زباله‌ها را جمع‌آوری و در مخزن خود ذخیره می‌کند. هدف این ربات جمع‌آوری مقدار بیشـــتری زباله از محیط است. چندین محدودیت اساسی دراین‌باره وجود دارند:

۱. انرژی ربات از طریق باتریِ قابل پرشدن (شارژ) تأمین می‌شود. بنابراین، ربات با بررسی مقدار انرژی باقی‌مانده، باید به‌گونه‌ای برنامه‌ریزی کند که قبل از تمام‌شدن کامل باتری، خود را به محل پرشدن (شارژ) برساند.

۲. در محیط موانعی مانند آتش قرار دارند که ربات در صورت برخورد با آن‌ها مشتعل می‌شود و از بین می‌رود.

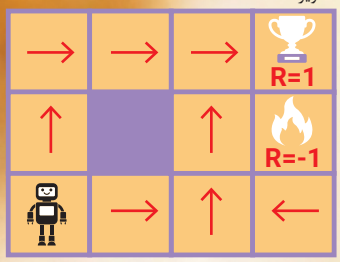
۳. ظرفیت و حجم مخزن نگهداری زباله در ربات محدود است. پس از پرشدن مخزن ربات، باید به محل تخلیهٔ زباله مراجعه و مخزن را تخلیه کرد.

۴. دنیای واقعی دنیایی غیرقطعی است. به این معنا که دقیقـــاً همان تصمیمی که می‌گیریم، به‌طور قطعی قابل اجرا نیست. به‌طور مثال، ربات تصمیم می‌گیرد به سمت بالا حرکت کند. اما به احتمال ۸۰ درصد به سمت بالا، ۵ درصد به ســـمت چپ، ۵ درصد به راست، و ۵ درصد به پایین می‌رود. به‌احتمال ۵ درصد هم در همان خانه باقی می‌ماند. این حرکت غیرقطعی از مشکلات مکانیکی ربات ناشی می‌شود.

۵. محیط پویاســـت و محل زباله، آتش و موانع در هر لحظه تغییر می‌کند.

به‌منظور ساده‌سازی موضوع، تنها چالش دوم را در نظر می‌گیریم و سعی می‌کنیم عمل بهینه برای هر حالت را پیدا کنیم. به این معنا که اگر ربات در یک حالت (خانه) باشد، کدام عمل از مجموعه اعمال گفته‌شده (بالا، راست، چپ و پایین) را باید انتخاب کند تا در آینده بیشـــترین پـــاداش را از محیط دریافت کنـــد. تصویر ۱، حالت‌های ممکن و میزان پاداش محیـــط را در صورت یافتن زباله و همچنین میزان جریمـــه را در صورت برخورد با آتش نشان می‌دهد. در خانه‌ای که آتش قرار دارد، مقدار جریمهٔ دریافتـــی برابر با منفی یک و مقـــدار پاداش دریافتی در حالتی که زباله وجود داشته باشد، برابر با یک خواهد بود. (تصویر ۱)

تصویر ۳



برای یافتن سیاست بهینه باید ارزش هر حالت (خانه) را بدانیم. ارزش هر حالت، میزان احتمال رسیدن به پاداش در آینده را مشخص می‌کند. برای محاسبهٔ ارزش حالت‌ها، از همسایگان حالت پاداش شروع می‌کنیم و مقدار ارزش آن‌ها را برابر با یک قرار می‌دهیم. در مثال شکل بالا، تنها یک همسایه با حالت پاداش داریم. از آنجا که این حالت به‌طور مستقیم ما را به حالت پاداش می‌رساند، بیشترین ارزش را خواهد داشت. سپس ارزش خانه‌های همسایه را با خانه‌ای که در مرحلهٔ قبل ارزش آن را محاسبه کردیم به دست می‌آوریم. این خانه‌ها، همسایگان دوگانهٔ حالت پاداش هســـتند. بنابراین، مقدار پاداش آن‌ها یک مقدار کاهش پیدا می‌کند. اگر فـــرض کنیم اندازهٔ این کاهش برابر با ۰٫۹ باشد، مقدار ارزش به‌صورت ضریبی از ۰٫۹ در خانه‌های کناری کاهش می‌یابد. این محاسبات تا تعیین ارزش همهٔ خانه‌ها ادامه می‌یابد. شـــکل شـــماره ۲ این محاسبات را مرحله به مرحله نشان می‌دهد.

در پایان، حرکت بهینه بر اساس ارزش حالت‌ها تعیین می‌شود، به‌گونه‌ای که در هر حالت مشخص می‌شود کدام عمل باید انجام شـــود. تصویر ۳ این سیاست را بر اساس ارزش‌های تعیین‌شـــده در تصویر ۲ نشان می‌دهد. وقتی ربات تصمیم به حرکت می‌گیرد، می‌تواند یکی از دو عمل «راست» یا «بالا» را انتخاب کند. سپس عمل مناسب را بر اساس سیاست، تا رسیدن به پاداش، انجام خواهد داد.