علی اکبری آلاشتی
استاد بابک فرهادی
شماره دانشجویی ۴۰۲۱۲۳۳۰۱۱۷۱۶۳

یادگیری تقویتی: یادگیری آن‌پالیسی و آف‌پالیسی

یادگیری تقویتی (Reinforcement Learning) یکی از شاخه‌های اصلی یادگیری ماشین است که در آن یک عامل (Agent) از طریق تعامل با محیط (Environment) و دریافت بازخورد به‌صورت پاداش یا جریمه، یاد می‌گیرد که چگونه تصمیم‌های بهینه‌ای اتخاذ کند. هدف اصلی در یادگیری تقویتی، یافتن سیاستی (Policy) است که مجموع پاداش‌های بلندمدت را بیشینه کند. سیاست، در واقع، یک تابع یا استراتژی است که مشخص می‌کند عامل در هر حالت از محیط چه عملی را انتخاب کند. در یادگیری تقویتی، دو رویکرد اصلی برای یادگیری سیاست وجود دارد: آن‌پالیسی (On-Policy) و آف‌پالیسی (Off-Policy). در ادامه، این دو رویکرد توضیح داده شده و تفاوت‌های آن‌ها بررسی می‌شود.

یادگیری **آن‌پالیسی** (On-Policy)
در روش‌های آن‌پالیسی، عاملی که سیاست را یاد می‌گیرد، از همان سیاستی که در حال بهبود آن است برای تولید داده‌ها (مانند انتخاب اقدامات و جمع‌آوری تجربه‌ها) استفاده می‌کند. به عبارت دیگر، سیاست مورد استفاده برای تعامل با محیط (Behavior Policy) همان سیاستی است که به‌روزرسانی و بهبود می‌یابد (Target Policy). این روش به عامل اجازه می‌دهد تا مستقیماً از تجربه‌های خود یاد بگیرد و سیاست را به‌تدریج بهبود دهد.

یکی از معروف‌ترین الگوریتم‌های آن‌پالیسی، **SARSA (State-Action-Reward-State-Action)** است. در SARSA، عامل در هر مرحله یک اقدام را بر اساس سیاست فعلی انتخاب می‌کند، پاداش و حالت بعدی را مشاهده می‌کند و سپس اقدام بعدی را نیز با استفاده از همان سیاست انتخاب می‌کند. این فرآیند به‌روزرسانی Q-valueها را بر اساس سیاست کنونی انجام می‌دهد. مزیت این روش این است که یادگیری به‌صورت مستقیم و پایدار انجام می‌شود، اما ممکن است به دلیل وابستگی به سیاست فعلی، کاوش محیط (Exploration) محدود شود و بهینه‌سازی کندتر انجام گیرد.

یادگیری **آف‌پالیسی** (Off-Policy)
در روش‌های آف‌پالیسی، سیاستی که برای تولید داده‌ها و تعامل با محیط استفاده می‌شود (Behavior Policy) متفاوت از سیاستی است که در حال یادگیری و بهبود است (Target Policy). این امکان را به عامل می‌دهد که از تجربه‌های جمع‌آوری‌شده توسط سیاست‌های دیگر (حتی سیاست‌های تصادفی یا کاوشگر) برای بهبود سیاست هدف استفاده کند. این ویژگی باعث می‌شود که روش‌های آف‌پالیسی معمولاً انعطاف‌پذیرتر و کارآمدتر باشند، به‌ویژه در سناریوهایی که داده‌های قبلی یا تجربه‌های جمع‌آوری‌شده توسط عامل‌های دیگر در دسترس است.

یکی از معروف‌ترین الگوریتم‌های آف‌پالیسی، Q-Learning است. در Q-Learning، عامل اقدام را بر اساس یک سیاست کاوشگر (مانند ε-greedy) انتخاب می‌کند، اما Q-valueها را با فرض انتخاب بهترین اقدام ممکن در حالت بعدی (سیاست حریصانه) به‌روزرسانی می‌کند.

این روش به عامل اجازه می‌دهد که به سمت سیاست بهینه همگرا شود، حتی اگر داده‌ها از سیاست‌های غیربهینه جمع‌آوری شده باشند. با این حال، آف‌پالیسی ممکن است به دلیل تفاوت بین سیاست‌های رفتار و هدف، ناپایداری‌هایی در یادگیری ایجاد کند.

تفاوت‌های کلیدی

۱. سیاست مورد استفاده:

در آن‌پالیسی، سیاست رفتار و هدف یکسان است، اما در آف‌پالیسی این دو متفاوت هستند.

۲. کارایی داده:

روش‌های آف‌پالیسی معمولاً از داده‌ها به‌صورت کارآمدتری استفاده می‌کنند، زیرا می‌توانند از تجربه‌های جمع‌آوری‌شده توسط سیاست‌های دیگر بهره ببرند.

۳. پایداری و سرعت یادگیری:

روش‌های آن‌پالیسی معمولاً پایدارتر هستند، اما ممکن است کندتر به سیاست بهینه برسند. در مقابل، روش‌های آف‌پالیسی می‌توانند سریع‌تر به سیاست بهینه همگرا شوند، اما ممکن است ناپایداری‌هایی داشته باشند.

۴. کاربردها:

روش‌های آن‌پالیسی برای محیط‌هایی که نیاز به کاوش مداوم دارند مناسب‌ترند، در حالی که روش‌های آف‌پالیسی برای استفاده از داده‌های موجود یا سناریوهای پیچیده‌تر مناسب هستند.

نتیجه‌گیری

یادگیری آن‌پالیسی و آف‌پالیسی دو رویکرد اساسی در یادگیری تقویتی هستند که هر یک مزایا و معایب خاص خود را دارند. انتخاب بین این دو روش به نوع مسئله، میزان داده‌های در دسترس، و نیاز به کاوش یا بهره‌برداری (Exploitation) بستگی دارد. الگوریتم‌هایی مانند SARSA و Q-Learning نمونه‌های کلاسیک این دو رویکرد هستند که به‌ترتیب برای یادگیری آن‌پالیسی و آف‌پالیسی استفاده می‌شوند. درک تفاوت‌های این روش‌ها به توسعه‌دهندگان کمک می‌کند تا الگوریتم مناسب را برای مسائل خاص انتخاب کنند.