به نام خدا

عنوان:

تکلیف دوازدهم درس یادگیری ماشین

استاد:

دکتر پدرام

دانشجو:

محمدعلی مجتهدسلیمانی

4033904504

تاریخ:

14/10/1403

سوال اول

نقش برنامه نویسی پویا در یادگیری مانند پیدا کردن کوتاه ترین مسیر از خانه تا دانشگاه است. ما میتوانیم راه های مختلفی را برای رسیدن به دانشگاه امتحان کنیم ولی ناکارآمد هستند. در عوض میتوانیم از یک نقشه (environment) استفاده بکنیم تا یک مسیری را برنامه ریزی بکنیم. ما مسئله را به قسمت های کوچک تری تقسیم میکنیم و سعی میکنیم در هر کدام از این قسمت های کوچک تر مسیر دنبال کوتاه ترین مسیر بگردیم و مجموع اینها باعث پیدا شدن کوچکترین مسیر بشوند. این ایده همان ایده برنامه نویسی پویا خواهد بود. حل کردن یک مسئله پیچیده با شکستن به قسمت های کوچکتر که با همدیگر همپوشانی دارند و حل کردن هر کدام از این قسمت های کوچکتر باعث ساختن راه حل مسئله اصلی ما میشود.

برنامه نویسی پویا و یادگیری تقویتی با همدیگر مرتبط هستند که میتوانیم در رویکرد های مختلف رد این ارتباط را ببینیم:

رویکرد model-based RL:

برنامه نویسی پویا به طور اصلی در این شیوه به کار میرود، جایی که agent به یک مدلی از محیط (environment) دسترسی دارد. این مدل شامل 2 ویژگی است:

1. Transition probability یا احتمال حرکت از یک state به state دیگر وقتی یک action را انجام میدهیم. 2. Reward function یا پاداشی که برای هر action در یک state دریافت میشود.

رویکرد محاسبه تابع value:

نقش اصلی برنامه نویسی پویا در یادگیری تقویتی محاسبه کردن مقدار بهینه تابع value است. این تابع ارزش (value) دراز مدت یک action را در یک حالت (state) خاص تخمین میزند. مقدار این تابع بیشینه پاداش جمع شده ممکن را که یک agent میتواند بدست بیاورد از هر حالت تخمین میزند. البته با فرض اینکه مدل دارد از سیاست (policy) بهینه تبعیت میکند. الگوریتم های پویایی مثل value iteration و policy iteration در این مرحله به کار میروند.

رویکرد value iteration:

در این رویکرد ما به صورت تکرار شونده مقدار تخمینی value function را بهبود میدهیم تا به مقدار بهینه همگرا شود. در این رویکرد از معادله Bellman استفاده میشود.

رویکرد policy iteration:

در این رویکرد ما 2 مرحله داریم: 1. Policy evaluation که در این مرحله ما مقدار تابع value را برای سیاست مد نظر محاسبه میکنیم. 2. Policy improvement که در این مرحله بر اساس مقدار فعلی تابع value یک سیاست بهتری را پیدا میکنیم که از سیاست فعلی بهتر باشد، این پیدا کردن سیاست بهتر توسط انتخاب یک action در هر حالت که ما را به بشینه پاداش مورد انتظار ما بر طبق تابع value فعلی برساند صورت میگیرد.

پیدا کردن سیاست بهینه:

وقتی مقدار بهینه تابع value محاسبه شد توسط هر کدام از رویکرد های بالا، ما سیاست بهینه را انتخاب میکنیم و agent به سادگی در هر حالت actionیی را انجام میدهد که پاداش مورد انتظار خودش را زیاد کند.

اگر چه که برنامه نویسی یک رویکرد model-based است، در رویکرد model-free هم از ایده های برنامه نویسی پویا استفاده میشود.