هوش مصنوعی

گزارش کار پروژه سوم

تاریخ : ۱۴۰۳/۰۲/۲۱۸

امیر محمد حکیمی

۴۰۱۳۱۰۱۱

گزارش 1:

زمانی میتوان همگرا بودن ان را اثبات کرد که γ کوچک تر از یک و همچنین k بزرگ باشد ولی اگر این شرایط نباشد یعنی γ برابر با یک باشد همگرایی ان تضمین نخواهد شد

گزارش2:

الگوریتم‌های تکرار ارزش و تکرار سیاست دو رویکرد مختلف برای حل مسائل کنترل بهینه در فضای زمان-عملگر هستند.

الگوریتم تکرار ارزش:(Value Iteration)

- در این الگوریتم، ما از تابع ارزش (value function) استفاده می‌کنیم تا بهبودی در سیاست‌ها ایجاد کنیم.
- مراحل اصلی عبارتند از:

۱. **Initialization**: مقدار اولیه تابع ارزش را تعیین می‌کنیم.

۲. **Backup (Bellman Backup)**: با استفاده از تابع ارزش، مقدار ارزش هر حالت را به‌روزرسانی می‌کنیم.

۳. **Policy Extraction**: سیاست بهینه را با استفاده از تابع ارزش به‌دست می‌آوریم.

- این الگوریتم به شرطی همگرا می‌شود که ضریب تخفیف (γ) کمتر از یک باشد و تعداد تکرارها (k) به اندازه کافی بزرگ باشد.

۲. **الگوریتم تکرار سیاست:(Policy Iteration)**

- در این الگوریتم، هدف ما بهبود سیاست اولیه و رسیدن به سیاست بهینه است.
- مراحل اصلی عبارتند از:

۱. **Policy Evaluation**: ارزیابی سیاست فعلی با استفاده از تابع ارزش.

۲. **Policy Improvement**: بهبود سیاست با تغییر عمل‌ها به‌صورت مستقیم.

- این الگوریتم نیز مانند Value Iteration با تعداد تکرارهای کافی همگرا می‌شود.

- مزیت این الگوریتم این است که ممکن است با تعداد تکرار کمتری نسبت به Value Iterationبه جواب برسیم، اما هر تکرار آن عملیات محاسباتی پیچیده‌تر از تکرار ارزش است. با این حال، در اکثر مواقع، تکرار سیاست سریع‌تر از تکرار ارزش است

گزارش 3:

تعادل بین exploration و exploitation در یادگیری تقریباً در همه مسائل کنترل بهینه مطرح است:

۱. **Exploration**:

- در مرحله اکتشاف، عامل سعی می‌کند حالت‌ها و عمل‌ها را به طور جامع بررسی کند.
- این به منظور کشف حالت‌های ناشناخته، تجربه اقدامات مختلف، و کاهش عدم قطعیت در محیط است.
- مقدار بالای epsilon در الگوریتم‌های مانند Q-learning یاε-greedy ، به عامل اجازه می‌دهد تا در ابتدای یادگیری اقدامات تصادفی انجام دهد و مسیرهای غیر بهینه را بررسی کند.

۲. **Exploitation**:

- در مرحله بهره‌برداری، عامل از دانش خود استفاده می‌کند تا بهینه‌ترین اقدامات را انتخاب کند.
- این به منظور بهبود کارایی و کسب پاداش بیشتر است.
- با کاهش مقدارepsilon ، عامل به تدریج از اکتشاف به بهره‌برداری تغییر می‌دهد و مسیر بهینه‌تر را دنبال می‌کند.

فرایند کاهش epsilon به‌عنوان یک روش تعادل‌دهی بین اکتشاف و بهره‌برداری، موجب پایداری فرایند یادگیری می‌شود و در نهایت به مسیر بهینه همگرا می‌شود.

گزارش 4:

الگوریتم Q-Learning یک الگوریتم **Off-Policy** است. به دیگر الگوریتم‌های Off-Policy مانند Q-Learning ، تخمین پاداش برای جفت‌های حالت-عمل بر اساس سیاست بهینه (به صورت خودخواه) انجام می‌شود و مستقل از اقدامات عامل است . عبارت دیگر، الگوریتم‌های Off-Policy تخمین‌های مقدار عمل بهینه را مستقل از سیاست مشخص می‌کنند .این الگوریتم‌ها قادر به به‌روزرسانی مقادیر تخمین‌زده‌شده با استفاده از اقدامات مصنوعی نیز هستند.

برای درک بهتر، بیایید به توضیحات زیر بپردازیم:

- **Policy-On**:
  - در الگوریتم‌های Policy-On ، تخمین‌گرها بر اساس سیاست فعلی عامل به‌روزرسانی می‌شوند.
  - به عبارت دیگر، مقدار تخمین‌زده‌شده برای هر حالت-عمل به توجه به اقداماتی است که عامل در سیاست فعلی انجام می‌دهد.
  - مثالی از الگوریتم‌های Policy-On ، SARSA است.
- **Policy-Off**:
  - در الگوریتم‌های Policy-Off ، تخمین‌گرها مستقل از سیاست فعلی عامل به‌روزرسانی می‌شوند.
  - به عبارت دیگر، مقدار تخمین‌زده‌شده برای هر حالت-عمل مستقل از اقدامات عامل است.
  - مثالی از الگوریتم‌های Policy-Off ، Q-Learning است.

با توجه به ماهیت Off-Policy الگوریتم Q-Learning ، می‌تواند از تجربیاتی که توسط هر سیاست دیگری جمع‌آوری شده‌اند، بهره ببرد .

گزارش 5:

الگوریتم Q-learning یکی از روش‌های یادگیری تقویتی است که از TD-learning و Monte Carloمشتق شده است.

۱. **Monte Carlo (MC):**

- درMC ، ما از تمام تجربه‌های یک قدمه (یعنی یک اپیزود کامل) برای به‌روزرسانی تخمین‌های ما استفاده می‌کنیم.
- مزیت: تخمین‌ها دقیق‌تر می‌شوند چون از تمام تجربه‌ها استفاده می‌کنیم.
- معایب: نیاز به انتظار تا پایان اپیزود داریم تا بتوانیم تخمین‌ها را به‌روز کنیم.

۲. **Temporal Difference (TD):**

- درTD ، ما تخمین‌ها را بر اساس تخمین‌های آینده به‌روز می‌کنیم، حتی قبل از اتمام اپیزود.
- مزیت: نیاز به انتظار نداریم و می‌توانیم تخمین‌ها را به‌روز کنیم.
- معایب: تخمین‌ها ممکن است ناپایدار باشند و به تغییرات در تجربه‌ها حساس باشند.

۳. **Q-learning:**

- Q-learningیک نوع TD(۰) است، به این معنی که از تخمین‌های آینده برای به‌روزرسانی Q-value استفاده می‌کند.
- مزیت: نیاز به انتظار نداریم و می‌توانیم تخمین‌ها را به‌روز کنیم.
- معایب: ممکن است در محیط‌های پیچیده‌تر ناپایدار باشد.

به طور کلی، MC برای محیط هایی که اپیزودهای طولانی تر هستند مناسب است، در حالی که TDو Q-learning برای محیط هایی با اپیزود های کوتاه تر مناسب تر هستند.