Методы RL

Подготовил: Нугаманов Эдуард

План рассказа:

1. Цель обучения
2. Функции ценности
3. Оптимальные функции ценности
4. Улучшение стратегии
5. Обобщенная итерация по стратегиям
6. Методы Монте-Карло
7. Обучение на основе временных различий

Цель обучения

<u>Максимизация ожидаемой выгоды</u>

Простейший случай:

$$R_t = r_{t+1} + r_{t+2} + \cdots + r_T$$

$r_{t+1}, r_{t+2}, \ldots,$ −последовательности вознаграждений после t

Приведенная выгода:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

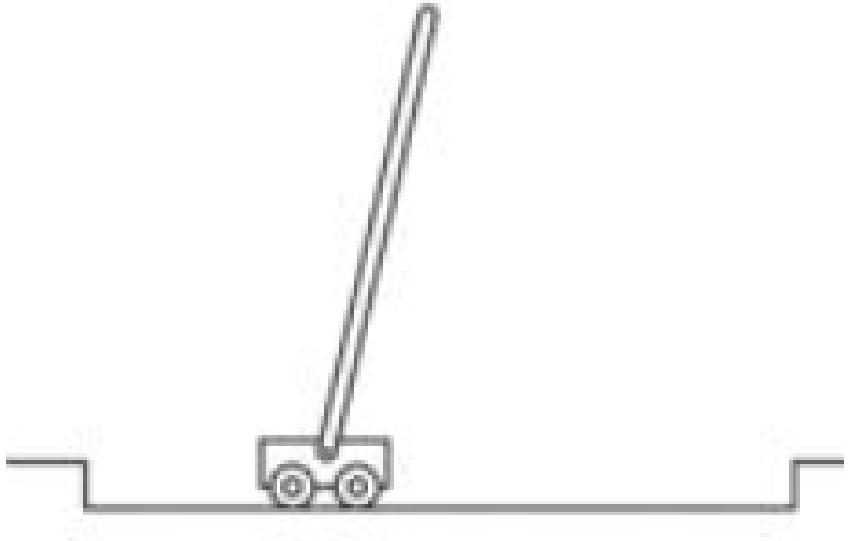
Пример: Балансировка стержня

Приложить силу к тележке так, чтобы стержень не падал.

Неудача – падение стержня или достижение границ площадки

- Эпизодический случай:
  - Вознаграждение +1 за шаг без неудачи
  - Выгода – число шагов до неудачи
- Непрерывный случай:
  - Вознаграждение -1 за неудачу, 0 все остальное время
  - Выгода зависит от $-\gamma^k$, где k – число шагов до неудачи

Максимизировать выгоду = сохранять баланс как можно дольше



Функции ценности

Оценивают *насколько хорошо* для агента находиться в данном состоянии *s* (осуществить данное действие *a* в данном состоянии *s*)

Ценность состояния для стратегии π:

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\}$$

Ценность действия для стратегии π:

$$Q^{\pi}(s,a) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\}$$

Оптимальные функции ценности

Частичный порядок на множестве стратегий:

$$\pi \geq \pi' \Leftrightarrow V^{\pi}(s) \geq V^{\pi'}(s)\ \forall s$$

Оптимальная функция ценности состояния:

$$V^{*}(s) = \max_{\pi} V^{\pi}(s)\ \forall s$$

Оптимальная функция ценности действия:

$$Q^{*}(s,a) = \max_{\pi} Q^{\pi}(s,a)\ \ \forall s, a \in \mathrm{A}(s)$$

Зная оптимальные функции, можно сделать оптимальные действия.

Улучшение стратегии

Для некоторого состояния s необходимо узнать нужно ли менять стратегию для отдельного действия $a \neq \pi(s)$

Теорема об улучшении стратегии:

$$Q_\pi\big(s, \pi'(s)\big) \geq V^\pi(s), \text{ тогда } \pi' \text{ не хуже, чем } \pi$$

Выбираем наилучшее действие согласно $Q^\pi(s, a)$:

$$\pi'(s) = \arg\max_a Q^\pi(s, a)$$

В случае ФМППР сходится к оптимальной стратегии.

Обобщенная итерация по стратегиям

Оцениваем стратегию и улучшаем.

Если процессы стабилизируются, то полученная функция ценности и стратегия – оптимальные.

Оценка стратегии методом Монте-Карло

Задача должна быть *эпизодической*

МК-метод первого посещения:

$$V^{\pi}(s) - \text{среднее значение выгод первых посещений } s$$

Все выгоды – i.i.d. оценки значений ценности $V^{\pi}(s)$, по ЗБЧ, последовательность их средних сходится к их мат. ожиданию

Оценка действия методом МК

Все аналогично оценке ценности состояния.

МК-метод первого посещения:

$Q^{\pi}(s,a) -$ среднее выгод при первом посещении s и действии a

Проблема поддерживающего изучения

Многие значимые пары состояние-действие никогда не посещены!

Если π – детерминированная стратегия, то получим значение выгоды только одного из действий для каждого состояния.

Необходимо найти ценности всех действий в каждом состоянии.

Изучающие старты

- Каждый эпизод начинается с пары состояние-действие.
- Каждая такая пара может быть выбрана как стартовая с ненулевой вероятностью.
- В пределе гарантирует посещение всех пар бесконечное число раз.
- На практике редкость, вместо используются *стохастические стратегии* с ненулевой вероятностью выбора всех действий.

ε-жадные стратегии

Гибкая стратегия: $\pi(s,a) > 0 \; \forall s,a$

Выбираем жадные действия, а с вероятностью ε – произвольные.

$$\frac{\varepsilon}{|A(s)|} - \text{минимальная вероятность выбрать нежад. действие}$$

Теорема об улучшении стратегии работает и здесь.

Управление по методу МК

Имеем функцию ценности действия – не нужна модель:

Методы на основе временных различий

- Методы не требуют предварительных знаний о модели поведения окружающей среды
- Обновление расчетных оценок, основанных на других оценках, не дожидаясь окончательного результата

- Предсказание
- Управление

Предсказание

В отличие от МК не надо ждать конца эпизода, достаточно следующего временного шага

TD(0):

$$V_{s_t} \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

При корректировке методом МК целью является выгода R_t, здесь:

$$r_{t+1} + \gamma V_t(s_{t+1})$$

SARSA

Корректировка ценности действия:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

Имеет место после каждого перехода из нетерминального состояния s_t. Если s_{t+1} - терминальное, то $Q(s_{t+1}, a_{t+1}) = 0$

Используем, например, ε-жадную стратегию.

Q-обучение

Одношаговое Q-обучение:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

Функция ценности действия аппроксимирует оптимальную, независимо от стратегии

При условиях: все пары корректируются и стохастической аппроксимации

$$Q_t \to Q^* \text{ почти наверное.}$$

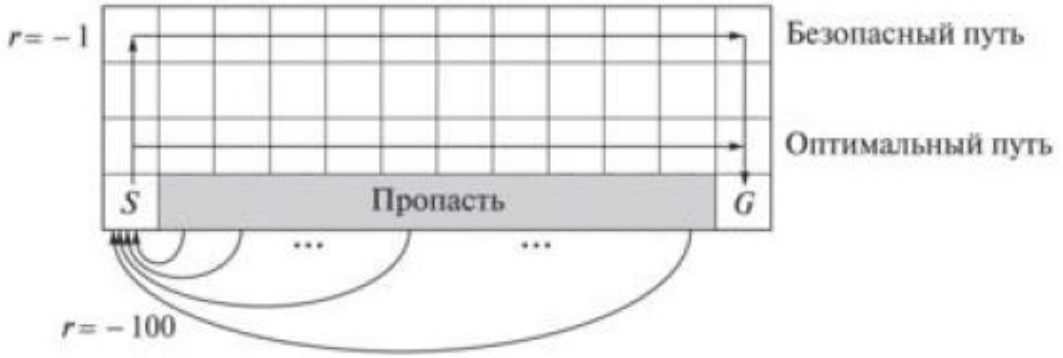
Сравнение: прогулка у пропасти

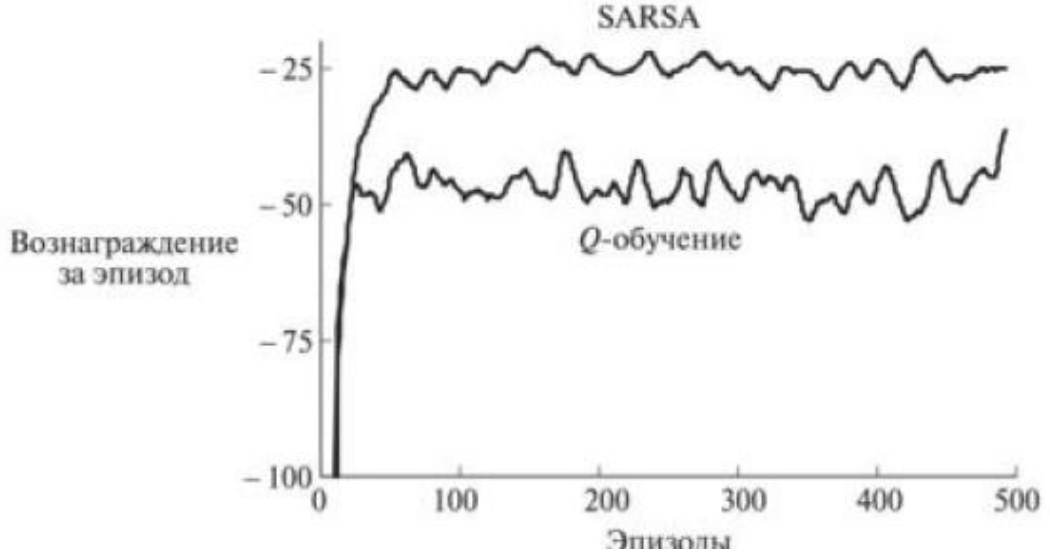
- ε-жадная стратегия

- Q-обучение старается найти оптимальную стратегию – ведет по краю пропасти.

- SARSA учитывает выбор действий и пытается найти более длинный и безопасный путь.

- Если уменьшать ε, оба метода сойдутся к оптимальной стратегии.





Список литературы

- Р.С. Саттон, Э.Г. Барто – Обучение с подкреплением