Методы решения задач обучения с подкреплением

Метод Монте-Карло

Метод Монте-Карло

- Что делать если у нас нет модели в виде переходных вероятностей и ожидаемых подкреплений?

- Мы можем использовать (усреднять) получаемые возвраты
  - Работает только для эпизодических задач
  - Может использоваться в процессе взаимодействия агента со средой
  - Может использовать модель, от которой требуется только генерировать возможные эпизоды

Оценка стратегий методом Монте-Карло

- Мы хотим определить ценность $V^{\pi}(s)$ состояния s для стратегии π.
- Имеются записи эпизодов, в которых агент управлялся π проходил через состояния s.
- Усредняем возвраты полученные после посещения состояния s.
  - Метод первого посещения
  - Метод любого посещения

    - оба сходятся в пределе если посещаются все состояния.

Алгоритм оценки стратегии методом Монте-Карло первого посещения

Вход: π – оцениваемая стратегия

Инициализация:

V(*s*) – произвольно, *Returns*(*s*) – пустой список для всех $s \in S$.

Повторять вечно

а) Генерировать эпизод используя π

б) Для всех состояний *s* в эпизоде

$R \leftarrow$Возврат после первого посещения *s*.

Добавить *R* в *Returns*(*s*)

$V(s) \leftarrow$Среднее значение (*Returns*(*s*))

Пример - 21

- Играют игрок и дилер
- Стоимость карт
  - Цифры – по своему номеру
  - Валет, дама, король – 10
  - Туз – 11 или 1
- В начале игры каждый получает по 2 карты, одна из карт дилера открыта
- Игрок может или брать по одной карте или остановится
- Дилер берёт карту если его сумма меньше 17, иначе останавливается
- Если кто-то набрал больше 21 он проигрывает
- Тот, кто набрал сумму ближе к 21 выигрывает

Пример - 21

- Эпизодический конечный МППР
- Подкрепление +1 (выигрыш), -1 (проигрыш) или 0 (ничья) в конце эпизода, остальные – 0
- Не используется дисконт $\gamma = 1$
- Действия – брать карту (hit) или остановится (stick)
- Состояния (200 штук):
  - Сумма карт игрока (12…21)
  - Открытая карта дилера (туз…10)
  - Есть ли у игрока туз, который можно считать как 11 (да/нет)

Пример - 21

- Рассмотрим стратегию «брать карту пока сумма не станет равна 20 или 21»
- Моделируем много игр и усредняем возвраты полученные после каждого из состояний
- Состояния в игре не повторяются, поэтому нет разницы между методом первого и любого посещений.

Пример - 21

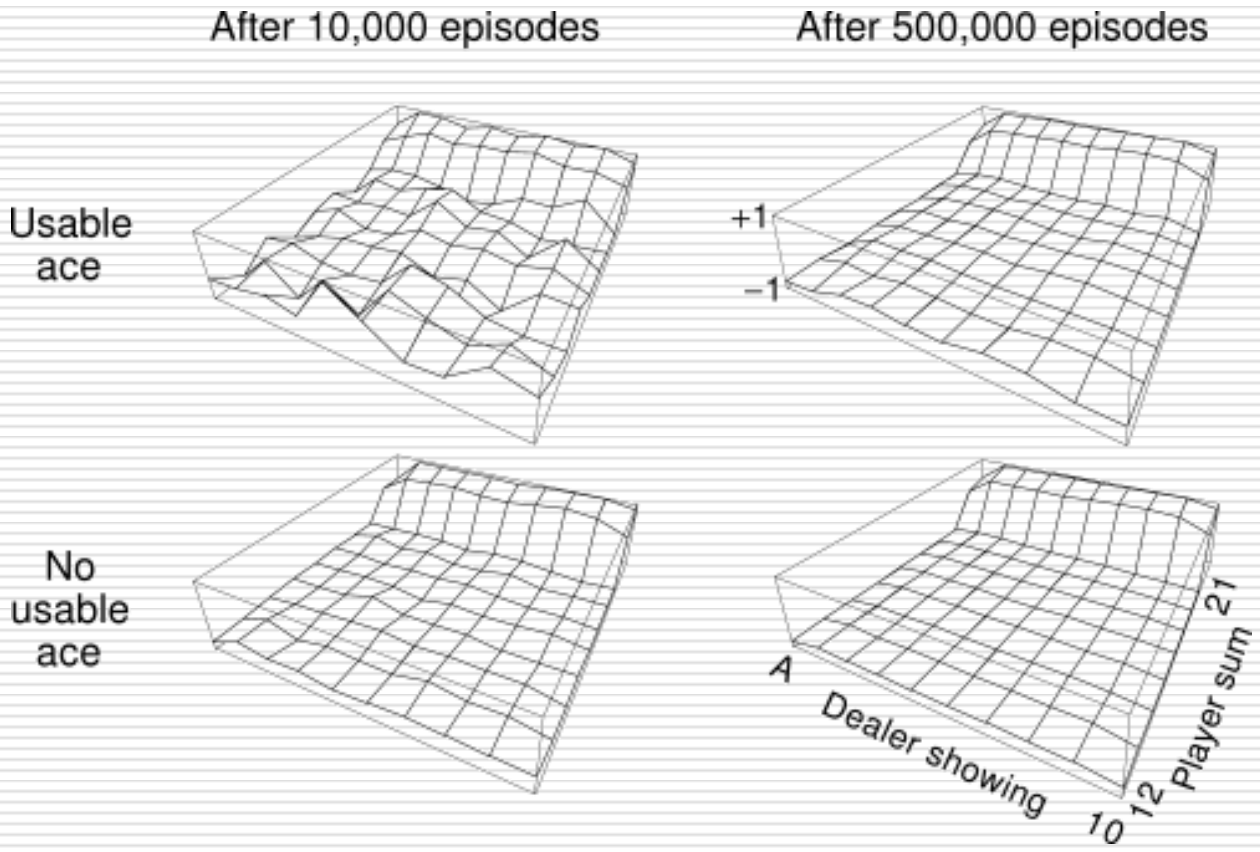
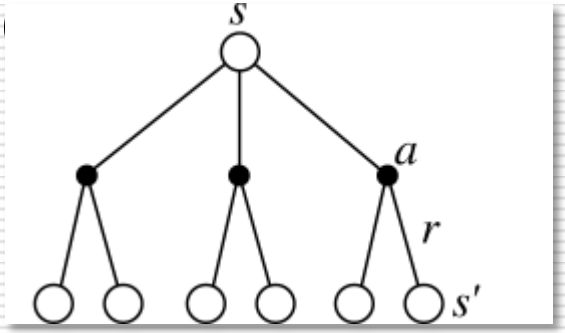

After 10,000 episodes
After 500,000 episodes
Usable ace
No usable ace
+1
−1
A
Dealer showing
10
12
Player sum
21

Диаграмма обновлений метода Монте-Карло

Метод Монте-Карло

Динамическое программирование



Пример – мыльная плёнка

- Какую форму примет мыльная плёнка, натянутая на изогнутую проволочную рамку?
- Сила, действующая со стороны соседей в каждой точке равна 0
  - Высота в каждой точке равна среднему значению высоты в ближайшей окрестности
  - Значения высоты для краевых точек равны высоте рамки

Пример – мыльная плёнка

- Разобьём поверхность сеткой
- Вариант 1
  - Фиксируем краевые точки по высоте рамки
  - В цикле заменяем высоту в каждой внутренней точке на среднюю высоту соседей
- Вариант 2
  - Берём внутреннюю точку и генерируем много случайных путей до границы
  - Высота в точке равна среднему значению достигнутых точек границы

Оценка стратегий методом Монте-Карло

- Хорошо
  - Работает на основе опыта
  - Вычисления для одних состояний не зависят от вычислений для других состояний
- Плохо
  - Не имея модели мы не можем на базе V строить жадную стратегию

    $\Rightarrow$ Нужно оценивать не V а Q.

Оценка действий методом Монте-Карло

- Аналогичным образом оцениваем $Q(s,a)$ на основе имеющегося опыта.

- Мы должны посещать все пары (s,a), т.е. для каждого состояния мы должны пробовать все действия.
  - Будем рассматривать эпизоды начинающиеся со всех возможных пар (s,a)
  - Случайные стратегии

Итерация стратегий методом Монте-Карло

- Используем Q, а не V:

$$\pi_0 \xrightarrow{o} Q^{\pi_0} \xrightarrow{y} \pi_1 \xrightarrow{o} Q^{\pi_1} \xrightarrow{y} \pi_2 \xrightarrow{o} \cdots \xrightarrow{y} \pi^* \xrightarrow{o} Q^*$$

- Жадная стратегия (улучшение):

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

- Сходится если:
  - Начинаем со всех возможных (*s*,*a*)
  - Вычисляем оценку наблюдая бесконечное число эпизодов

Итерация стратегий методом Монте-Карло

- Что делать с бесконечным числом эпизодов?
  - Если мы хотим вычислять Q^{π} на каждом шаге: можем оценивать возможное отклонение и ждать пока оно не станет слишком малым.
  - Можем не ждать пока завершится вычисление Q^{π}, а переходить к улучшению после некоторого числа вычислений
    - Для методов Монте-Карло естественно проводить вычисления после окончания эпизода

Алгоритм Монте-Карло с произвольным началом

Инициализация:

$Q(s,a)$,$\pi(s)$ – произвольно, *Returns*(s,a) – пустой список для всех $s \in S, a \in A(s)$.

Повторять вечно

а) Генерировать эпизод используя π и все возможные варианты (s,a) в качестве начала

б) Для всех пар (s,a) в эпизоде

$R \leftarrow$Возврат после первого посещения (s,a).

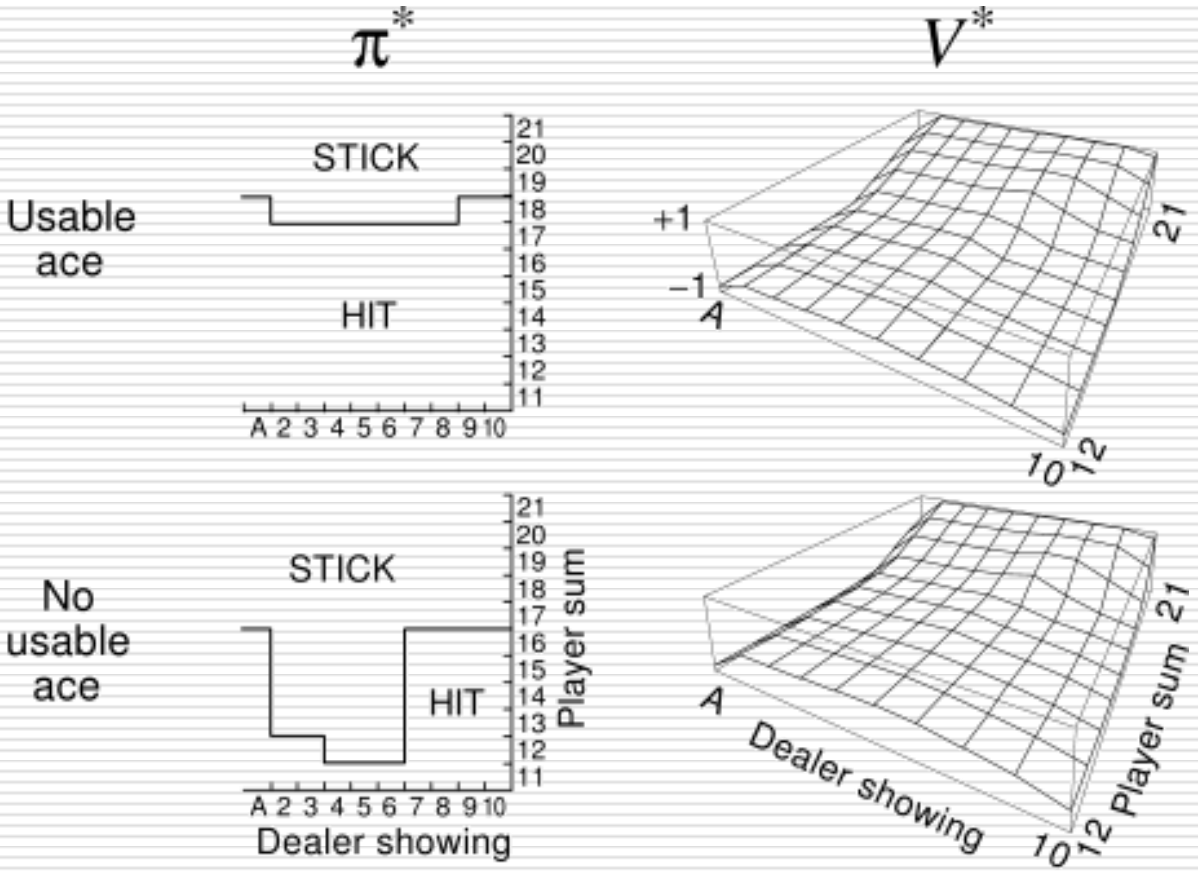
Добавить R в *Returns*(s,a)

$Q(s,a) \leftarrow$Среднее значение (*Returns*(s,a))

в) Для всех состояний s в эпизоде

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

Пример – как играть в 21



Управление методом Монте-Карло

- Как отказаться от допущения произвольного начала эпизода?
- Агент должен выбирать все возможные действия в процессе:
  - Управляем и оцениваем одну стратегию
    - Стратегия должна быть случайной
  - Управляем и оцениваем разные стратегии
    - Можем оценить не случайную стратегию

Управляем методом Монте-Карло следуя оцениваемой стратегии

- Агент должен пробовать все возможные варианты действий $\Rightarrow$ стратегия должна быть мягкой, т.е. $\pi(s,a)>0$ для всех s,a.
- ε-мягкие стратегии: $\pi(s,a)\geq\varepsilon/|A(s)|$
- ε-жадные $\pi(s,a)=\varepsilon/|A(s)|$ для не жадных действий, остальное – для жадных.

Управляем методом Монте-Карло следуя оцениваемой стратегии

- Используем идею обобщенной итерации стратегий
- Мы не можем делать стратегию жадной, но можем смещать её в направлении жадной – используем ε-жадные стратегии.
- Для любой ε-мягкой стратегии π, любая ε-жадная относительно Q^{π} стратегия гарантированно не хуже π.

Итерация ε-жадных стратегий

- Пусть π' - ε-жадная относительно Q^π стратегия

$$Q^\pi\big(s,\pi'(s)\big) = \sum_a \pi'(s,a)Q^\pi(s,a)$$

$$= \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) + (1-\varepsilon)\max_a Q^\pi(s,a)$$

$$\geq \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) + (1-\varepsilon)\sum_a \frac{\pi(s,a) - \frac{\varepsilon}{|\mathcal{A}(s)|}}{1-\varepsilon} Q^\pi(s,a)$$

$$= \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) - \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) + \sum_a \pi(s,a)Q^\pi(s,a)$$

$$= V^\pi(s)$$

Следовательно $\pi' \geq \pi$.

Итерация ε-жадных стратегий

Покажем, что $\pi' = \pi$ только когда обе стратегии являются оптимальными, среди ε-жадных.

Рассмотрим среду, которая с вероятностью $1 - \varepsilon$ ведёт себя как исходная, а с вероятностью ε - так, как будто было случайно выбрано другое действие.

Лучший результат, который может быть достигнут в такой среде соответствует лучшему результату который можно достичь в исходной среде применяя ε-жадные стратегии.

Пусть $\tilde{V}^*$ и $\tilde{Q}^*$ - оптимальные функции ценности для новой среды.

Тогда стратегия будет оптимальна среди ε-жадных тогда и только тогда, когда $V^\pi = \tilde{V}^*$.

Итерация ε-жадных стратегий

$$\begin{aligned}\widetilde{V}^*(s) &= (1-\varepsilon)\max_a \widetilde{Q}^*(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a \widetilde{Q}^*(s,a) \\ &= (1-\varepsilon)\max_a \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma \widetilde{V}^*(s')\right] \\ &\quad + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a\sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma \widetilde{V}^*(s')\right].\end{aligned}$$

$$\begin{aligned}V^\pi(s) &= (1-\varepsilon)\max_a Q^\pi(s,a) + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a Q^\pi(s,a) \\ &= (1-\varepsilon)\max_a \sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right] \\ &\quad + \frac{\varepsilon}{|\mathcal{A}(s)|}\sum_a\sum_{s'} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right].\end{aligned}$$

Так как $\tilde{V}^*$ - единственное решение, то $V^\pi = \tilde{V}^*$.

Алгоритм оценки текущей стратегии методом Монте-Карло.

Инициализация:

$Q(s,a)$ – произвольно, $\pi(s)$ – любая ε-жадная стратегия, $Returns(s,a)$ – пустой список.

Повторять вечно

а) Генерировать эпизод используя π

б) Для всех пар (s,a) в эпизоде

$R \leftarrow$ Возврат после первого посещения (s,a).

Добавить R в $Returns(s,a)$

$Q(s,a) \leftarrow$ Среднее значение $(Returns(s,a))$

в) Для всех состояний s в эпизоде

$a^* \leftarrow \arg\max_a Q(s,a)$

Для всех $a \in A(s)$

Если $a=a^*$ то $\pi(s) \leftarrow 1-\varepsilon+\varepsilon/|A(s)|$

иначе $\pi(s) \leftarrow \varepsilon/|A(s)|$

Следуем одной стратегии, а оцениваем другую

- Имеющийся алгоритм может находить только ε-жадные стратегии.
- Оценивать оптимальную (жадную) стратегию мы можем только если будем управлять по другой стратегии.
- Можно ли получить функцию ценности для стратегии π, следуя другой стратегии π'?
  - Да, если из $\pi(s,a)>0 \Rightarrow \pi'(s,a)>0$

Следуем одной стратегии, а оцениваем другую

- В эпизоде, полученным действуя по стратегии π', рассмотрим последовательность состояний после *i*-го посещения состояния *s*.
- Пусть $p_i(s)$ и $p_i'(s)$ – вероятности этой последовательности в случае если мы действовали по стратегии π и π' соответственно.
- Тогда

$$V(s) = \frac{\sum_{i=1}^{n_s} \frac{p_i(s)}{p_i'(s)} R_i(s)}{\sum_{i=1}^{n_s} \frac{p_i(s)}{p_i'(s)}}.$$

Следуем одной стратегии, а оцениваем другую

- Итак, нужно найти

$$V(s) = \frac{\sum_{i=1}^{n_s} \frac{p_i(s)}{p'_i(s)} R_i(s)}{\sum_{i=1}^{n_s} \frac{p_i(s)}{p'_i(s)}}.$$

- Пусть $T_i(s)$ – время завершения i-го эпизода затрагивающего s. Тогда

$$p_i(s_t) = \prod_{k=t}^{T_i(s)-1} \pi(s_k, a_k) \mathcal{P}^{a_k}_{s_k s_{k+1}}$$

$$\frac{p_i(s_t)}{p'_i(s_t)} = \frac{\prod_{k=t}^{T_i(s)-1} \pi(s_k, a_k) \mathcal{P}^{a_k}_{s_k s_{k+1}}}{\prod_{k=t}^{T_i(s)-1} \pi'(s_k, a_k) \mathcal{P}^{a_k}_{s_k s_{k+1}}} = \prod_{k=t}^{T_i(s)-1} \frac{\pi(s_k, a_k)}{\pi'(s_k, a_k)}.$$

Алгоритм оценки одной стратегии, следуя другой

Инициализация:

$Q(s,a) \leftarrow$ произвольно

$N(s,a) \leftarrow 0$; числитель

$D(s,a) \leftarrow 0$; знаменатель

$\pi \leftarrow$ произвольная детерминированная стратегия

Повторять вечно:

а) Генерировать эпизод, используя ε-жадную стратегию π':

$s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_{T-1}, a_{T-1}, r_T, s_T$

б) τ- последний шаг когда $a_\tau \neq (s_\tau)$.

в) Для всех пар (s,a), появлявшихся в эпизоде после τ:

$t \leftarrow$время первого после τ посещения (s,a);

$w \leftarrow \prod_{k=t+1}^{T-1} \frac{1}{\pi'(s_k,a_k)}$;

$N(s,a) \leftarrow N(s,a) + wR_t$;

$D(s,a) \leftarrow D(s,a) + w$;

$Q(s,a) \leftarrow \frac{N(s,a)}{D(s,a)}$.

г) Для всех $s \in S$

$\pi(s) \leftarrow \arg max_a Q(s,a)$

Инкрементальное вычисление оценок

- Нам нужно вычислять взвешенное среднее вида

$$V_n = \frac{\sum_{k=1}^{n} w_k R_k}{\sum_{k=1}^{n} w_k}.$$

- Мы можем делать это инкрементальным образом, используя правило обновления

$$V_{n+1} = V_n + \frac{w_{n+1}}{W_{n+1}}\left[R_{n+1} - V_n\right]$$

где $W_{n+1} = W_n + w_{n+1}$ - сумма весов, $W_0 = 0$

Метод Монте-Карло

+ Не требует модели.
+ Обучается при непосредственном взаимодействии со средой.
+ Может использовать модель, дающую примеры эпизодов.
+ Просто работать с подмножеством состояний.
+ Мало подвержены проблемам при отклонении от Марковских состояний.

- Необходимо пробовать все действия
- Работают только для эпизодических задач
- «Делают выводы» только в конце эпизода