Задача о многоруком бандите

Формулировка задачи

Есть n игровых автоматов. Дёргая ручку $i = 1, ..., n$ мы каждый раз с вероятностью p_i получаем 1 рубль, а с вероятностью $1 - p_i$ не получаем ничего. Разрешается сделать $N >> 1$ шагов, на каждом шаге можно дёргать только одну ручку. Сами p_i априорно не известны. Но мы знаем, что $p_i \sim U[0, 1]$. Нужно найти стратегию выбора ручек такую, чтобы максимизировать доход.

Уравнение Вальда-Беллмана

Теперь наша задача такая: нужно сопоставить описание процесса выбора ручек с управляемым марковским процессом и получить соответствующее уравнение Вальда-Беллмана.

Определим пространства состояний: w-win, l-lose.

$$s = (w_1, l_1; ...; w_n, l_n)$$

$$\sum_{i=1}^{n}(w_i + l_i) = k \leq N$$

,

k - номер шага, w_i - то, сколько выигрышей было связано с i-ой ручкой к k-ому шагу, l_i - сколько неудач принесла i-ая ручка к шагу k.

Стратегия - это выбор на каждом шаге одной из ручек $a(s) \in \{1, ..., n\}$.

При этом мы ничего не знаем про вероятности переходов из одного состояния в другое:

$$(w_1, l_1; ..., w_i, l_i; ...; w_n, l_n) \to (w_1, l_1; ...; w_i + 1, l_i; ...; w_n, l_n)$$

$$(w_1, l_1; ..., w_i, l_i; ...; w_n, l_n) \to (w_1, l_1; ...; w_i, l_i + 1; ...; w_n, l_n)$$

Пусть

$$\rho_i^w(w_i, l_i) = P(w_i \leftarrow w_i + 1)$$
$$\rho_i^l(w_i, l_i) = P(l_i \leftarrow l_i + 1)$$

Функция вознаграждения определяется так: если выбрана ручка i (с её историей), то вознаграждение равно 1 рубль с вероятностью $\rho_i^w(w_i, l_i)$, и 0 рублей с вероятностью $\rho_i^l(w_i, l_i)$.

Слово об управляемых марковских процессах

Рассмотрим марковскую систему со временем $(t = 0, 1, 2, ..., 3)$ претерпевает случайные изменения. Будем считать, что всё время система может находиться лишь в конечном числе состояний S.

На каждом шаге система находится в одном из этих состояний, и на каждом шаге мы вольны выбирать исходя из своей стратегии в какое состояние из S перейти.

Исходя из того, в каком состоянии $s \in S$ находится система в текущий момент t и какое действие было выбрано алгоритмом $a \in A$, можно определить, с какой вероятностью система окажется в следующий момент времени $t + 1$ в состоянии $s' \in S$. Пусть это $p(s, a; s')$.

$$\sum_{s' \in S} p(s, a; s') = 1$$

В каждый момент времени мы получаем вознаграждение $r(s, a)$

Цель - получить максимальное итоговое вознаграждение:

$$V^*(s) = \max_a \mathbb{E} \sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)), \quad s_0 = s$$

$\gamma^t \in (0, 1]$ - этот параметр отвечает за то, как обесцениваются деньги.

$$V^*(s) = \max_a \mathbb{E} \sum_{t=0}^{\infty} \gamma^t r(s_t, a(s_t)) = \max_a \mathbb{E} \left(r(s_0, a(s_0)) + \gamma \sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a(s_{t+1})) \right) =$$

$$= \max_a (R(s_0, a) + \gamma \mathbb{E}_{s_1} V^*(s_1) = \max_a \left(R(s_0, a) + \gamma \sum_{s_1} p(s, a; s') V^*(s_1) \right)$$

В нашей задаче уравнение Вальда-Беллмана будет иметь такой вид:

$$V^*(w_1, l_1; ...; w_n, l_n) =$$

$$= \max_{i=1,\dots,n} \mathbb{E}_{\rho_i^w, \rho_i^l} \left(\rho_i^w(w_i, l_i) \cdot (1 + \gamma V^*(w_1, l_1; ...; w_i + 1, l_i; ...; w_n, l_n)) + \rho_i^l(w_i, l_i) \gamma V^*(w_1, l_1; ...; w_i, l_i + 1; ...; w_n, l_n) | (\right.$$

$$= \max_{i=1,\dots,n} \left(\frac{w_i + 1}{w_i + l_i + 2} (1 + \gamma V^*(w_1, l_1; ...; w_i + 1, l_i; ...; w_n, l_n)) + \frac{l_i + 1}{w_i + l_i + 2} \gamma V^*(w_1, l_1; ...; w_i, l_i + 1; ...; w_n, l_n) \right)$$

Решение такого уравнения дорогое экспоненциально.

Упростим задачу

Будем считать $\gamma < 1$ и $N \to \infty$. И рассмотрим случай, когда всего 2 ручки, вероятность на одной из которых известна и равна p.

Выигрыш: $V^*(w, l; p)$

Тогда уравнение Вальда-Беллмана будет такое:

$$V^*(w, l; p) = \max \left(\frac{p}{1 - \gamma}, \frac{w + 1}{w + l + 2} [1 + \gamma V^*(w + 1, l; p)] + \frac{l + 1}{w + l + 2} \gamma V^*(w, l + 1; p) \right)$$

При $w + l >> 1$ $V^*(w, l; p)$ недалеко от $(1 - \gamma)^{-1} \max \left(p, \dfrac{w}{w + l} \right)$

Определим **индекс Гиттинса** ручки с историей (w, l):

$$V^*(w, l; p) = \frac{p}{1 - \gamma}$$

Индекс Гиттинса - это решение этого уравнения, которое мы обозначим за $p_\gamma(w, l)$.

При $w + l \to \infty$ имеем $p_\gamma(w, l) \to (1 - \gamma)^{-1} \dfrac{w}{w + l}$.

Максимум $V^*(w_i, l_i; p)$ будет достигаться на той ручке, у которой наибольший индекс $p_\gamma(w_i, l_i)$. Тогда мы можем решить эту задачу, если знаем $p_\gamma(w, l)$.

Проблема отсутствия информации об управляемом марковском процессе, отвечающем задаче о многоруких бандитах, была решена в итоге с помощью индексов Гиттинса за счет байесовского подхода – задания априорного распределения. В общем случае возможность использовать байесовский подход в исследовании управляемых марковских процессов предполагает наличие вероятностной модели для переходных вероятностей и вознаграждений, зависящей от неизвестных параметров. Сам факт наличия такой параметрической модели далеко не всегда имеет место. Но еще более специальное предположение, которое мы существенным образом использовали, – это некоторая симметричность постановки задачи, позволившая искать решение в виде индексной стратегии. Естественно, возникает вопрос: что же делать в общем случае, когда ничего не известно и все, что можно делать, это только наблюдать за процессом и обучаться?

Для ответа на поставленный вопрос вернемся к достаточно общей модели управляемого марковского процесса с конечным числом состояний и стратегий (действий). Будем использовать общие обозначения, введенные в разделе В.3. Только сделаем для удобства обозначений, одно небольшое уточнение: будем считать, что функция вознаграждения всецело определяется текущим состоянием, выбранной стратегией и состоянием, в которое перейдет процесс на следующем шаге. Таким образом, предполагается, что случайность в функции вознаграждения всецело определяется состоянием, в которое переходит процесс. Заметим, что и примеры с разборчивой невестой и с многорукими бандитами подходят под это предположение. В сделанных предположениях уравнение Вальда–Беллмана будет иметь вид

$$V^*(s) = \max_a \sum_{s' \in S} p(s,a;s') \left(r(s,a;s') + \gamma V^*(s')\right)$$

Введём Q-функцию

$$Q(s,a) = \sum_{s' \in S} p(s,a;s') \left(r(s,a;s') + \gamma V^*(s')\right)$$

$$V^*(s) = \max_a Q(s,a)$$

Тогда Q-функция должна удовлетворять Q-уравнению:

$$Q(s,a) = \sum_{s' \in S} p(s,a;s') \Big(r(s,a;s') + \gamma \max_{a'} Q(s',a')\Big)$$

Это уравнение можно решить методом последовательных итераций. Можно смотреть на $Q = \{Q(s,a)\}_{s\in S, a\in A}$ как на вектор. Тогда надо решить уравнение

$$Q = H(Q)$$

где H - сжимающий оператор, то есть он сопоставляет каждой точке точку, которая не менее, чем в γ раз ближе к началу координат.

Метод последовательных итераций будет иметь вид

$$Q_{t+1} = H(Q_t)$$

Тем не менее мы не очень владеем информацией о функциях $r(s,a;s')$ и $p(s,a;s')$

Суть Q-обучения - заменить невычислимое $H(Q_t)$ (так как мы не знаем H) на его вычислимую несмещённую оценку:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t(s,a) \Big(r(s,a;s'(s,a)) + \gamma \max_{a'} Q_t(s'(s,a),a') - Q_t(s,a)\Big)$$

Здесь $s'(s,a)$ - положение процесса на шаге $t+1$, если на шаге t процесс был в состоянии s и было выбрано действие a. Если на шаге t процесс находился в состоянии s и было выбрано действие a, то $0 < \alpha_t(s,a) \leq 1$, иначе $\alpha_t(s,a) = 0$.

Здесь уже мы знаем как вычислить функцию, так как на текущем шаге $t+1$ мы знаем результат шага t, также мы знаем, какое вознаграждение получаем, поскольку если мы переходим из состояния s в $s'(s,a)$, то мы можем просто наблюдать, какое вознаграждение получаем, а если не переходим в $s'(s,a)$, то есть не наблюдаем этот переход, то $\alpha_t(s,a) = 0$, так что в этом случае нет необходимости считать вознаграждение.

Заметим, что если наша стратегия $a(s)$ приводит к тому, что с вероятностью 1 каждая пара (s,a) будет бесконечное число раз встречаться на бесконечном горизонте наблюдения, то если будет выполнено условие сжимаемости

$$\sum_{t=0}^{\infty} \alpha_t(s,a) = \infty$$

$$\sum_{t=0}^{\infty} \alpha_t^2(s,a) < \infty$$

тогда процесс $Q_{t+1} = H(Q_t)$ сойдётся:

$$\lim_{t \to \infty} Q_t(s,a) = Q(s,a)$$

$$V^*(s) = \max_a Q(s,a)$$

Мы пришли к тому, что проделав большое число шагов, мы можем определить оптимальную стратегию, не зная никакую информацию про управляемый марковский процесс.

Правда проблема в том, что мы не знаем, сколько шагов нужно сделать, чтобы считать, что можно закончить обучение. Мы не знаем, в какой момент можно переходить на стратегию

$$a_t(s) = \arg\max_a Q_t(s,a)$$

со стратегии, которой придерживались сначала, которую мы сначала выбрали для наискорейшей сходимости процесса.

То есть идея в том, что до какого-то момента мы исследуем все ручки, и только после того как все ручки были проверены достаточное число раз, мы выбираем наилучшую из них.

Возвращаемся к исходной постановке

Будем считать $\gamma = 1$. Пусть нам дали N шагов и предположим, что мы знаем оптимальную ручку (у неё успех $p_{\max}$. Тогда можем получить ожидаемое вознаграждение $p_{\max}N$. Оказывается, что если мы ничего о ручках не знаем, то мы не сможем получить ожидаемое вознаграждение больше, чем

$$p_{\max}N - 0.05\sqrt{Nn}$$

Как можно приблизиться к такой оценке? Алгоритм $Exp3$ обеспечивает вознаграждение не меньше чем

$$p_{\max}N - 2\sqrt{Nn\ln n}$$

Exp3 алгоритм

На k-ом шаге выбираем ручку i с вероятностью

$$p_i^k = \frac{\exp(\eta_N R_i^k)}{\sum_{j=1}^n \exp(\eta_N R_j^k)}$$

$$\eta_N = \sqrt{\frac{2\ln n}{Nn}}$$