Обучение с подкреплением

Шевчук П.

Факультет компьютерных наук
Высшая школа экономики

20 октября 2016 г.

Задача о многоруком бандите

Формулировка

Среда содержит

- A - множество действий
- p(a, r) - распределение наград для каждого действия

Цель - найти стратегию, максимизирующую прибыль.
$Q_t(a) = \frac{\sum r_i[a_i=a]}{\sum[a_i=a]}$ - средняя награда за действие a

$\lim_{t\to\infty} E[Q_t(a)] = Q^*(a)$ - ценность действия.
Играем со средой по следующему алгоритму:

- Инициализируется стратегия
- На каждом шаге:
- агент выбирает действие на основе стратегии
- среда возвращает reward
- агент корректирует стратегию

Задача о многоруком бандите

Жадная стратегия

Жадная стратегия заключается в том, что мы выбираем действие с максимальной оценкой ценности, то есть:

$$A_t = \arg\max_{a \in A} (Q_t(a))$$

Недостаток этой стратегии в том, что мы почти не исследуем среду.
Эвристика - используем ε-жадную стратегию, то есть будем выбирать действие согласно жадной стратегии с вероятностью $1 - \varepsilon$ и случайное - с вероятностью ε.
Эвристика: ε можно уменьшать со временем.

Метод UCB заключается в том, что мы выбираем действие с максимальной верхней оценкой ценности, а именно

$$A_t = \arg\max_{a \in A} \left(Q_t(a) + \delta \sqrt{\frac{2 \ln t}{k_t(a)}} \right)$$

Интерпретация:

- Чем меньше $k_i(a)$, тем менее стратегия исследована, соответственно, вероятность должна быть больше
- δ - параметр, чем он больше, тем стратегия более исследовательская

Эвристика: δ можно уменьшать со временем.

Задача о многоруком бандите

Метод Softmax (распределение Больцмана)

Мягкий вариант компромисса между исследованием и применением: выбираем действие случайно из распределения, в котором вероятность равна:

$$\pi_t(a) = \frac{e^{\frac{Q_t(a)}{\tau}}}{\sum\limits_{b \in A} e^{\frac{Q_t(b)}{\tau}}}$$

τ - параметр *температуры*

- При $\tau \to 0$ стратегия стремится к жадной
- При $\tau \to \infty$ стратегия стремится к равномерной, то есть полностью исследовательской

Задача о многоруком бандите

Оценка методов

Генерируется 2000 задач, в каждой из которых:

- $|A| = 10$
- $p_a(r) = N(Q^*(a), 1)$
- $Q^*(a)$ выбирается случайно из $N(0, 1)$.

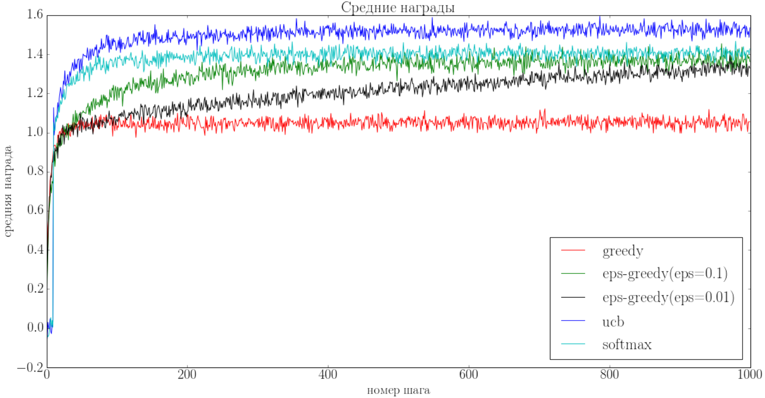
Строятся графики:

- Средняя награда
- Процент оптимальных действий

в зависимости от t, усредненное по всем 2000 задачам.

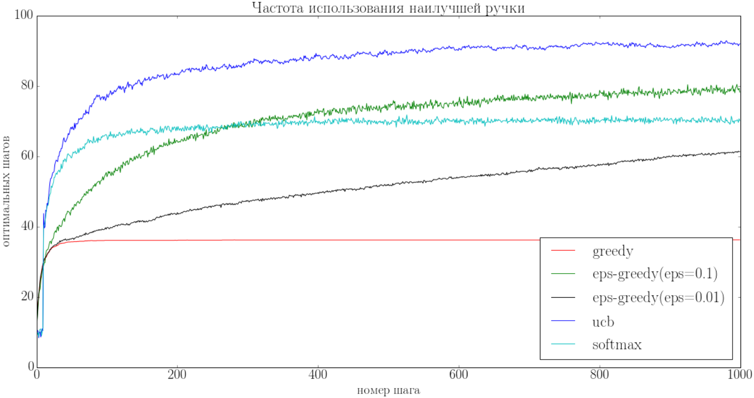
Задача о многоруком бандите

Графики



Задача о многоруком бандите

Графики



В реальности часто встречаются ситуации, когда среда изменяется, однако мы используем в алгоритмах данные о всех предыдущих попытках.
Решение: можно использовать вместо обычного среднего скользящее.

Задача о многоруком бандите

Метод сравнения с подкреплением

Идея: использовать не сами значения премий, а их разности с эталонным.

Действительно, мы можем сказать, большая или маленькая полученная награда, только сравнив с их общим уровнем.

- $\bar{r}_{t+1} = \bar{r}_t + \alpha(r_t - \bar{r}_t)$ - эталонная награда, скользящее среднее наград
- $p_{t+1}(a_t) = p_t(a_t) + \beta(r_t - \bar{r}_t)$ - предпочтения действий
- Используем Softmax с $p_t(a)$ вместо $Q_t(a)$

Эвристика: оптимистично завышенно $\bar{r}_0$ делает стратегию более исследовательской.

Экспериментальный факт: сравнения с подкреплением сходятся быстрее, чем ε-жадная стратегия.

Общая задача

Формулировка

Добавляется множество состояний, для каждой пары (действие, состояние) есть распределение наград и распределение состояний. Игра выглядит так:

- Инициализируется стратегия
- На каждом шаге:
- агент выбирает действие на основе стратегии и предыдущих состояний.
- среда возвращает награду и новое состояние
- агент корректирует стратегию

Общая задача

Приведённая выгода

Чтобы что-то оптимизировать нужно выбрать целевую величину:

$$R_t = r_{t+1} + r_{t+2} * \gamma + r_{t+2} * \gamma^2 + ... = \sum_{i=1}^{+\infty} r_{t+i}\gamma^{i-1}$$

$0 < \gamma < 1$ - коэффициент приведения.
Аналогия: net present value и дефлятор.

Общая задача

Приведённая выгода

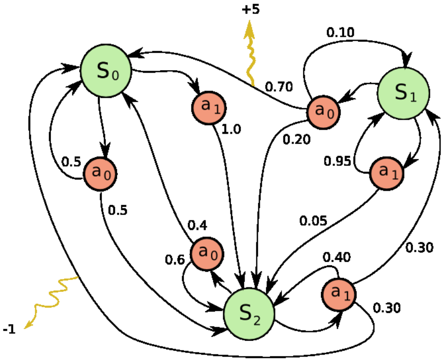
Чтобы нормально обрабатывать случай с конечной последовательностью добавим поглощающее состояние:

$$r_{\tau+1} = r_{\tau+2} = r_{\tau+3} = \ldots = 0$$

Общая задача

МППР

Это *Марковский процесс принятия решений (МППР)*, что значит, что следующее состояние зависит только от одного предыдущего, а не от всех.



Общая задача

Ценность состояния

Довольно полезно иметь какую-то оценку того, насколько хорошо состояние, в котором мы находимся.

$$V_\pi(s) = E[R_t | s_t = s] = E[\sum_{i=1}^{+\infty} r_{t+i} \gamma^{i-1} | s_t = s]$$

Общая задача

Ценность действия

Ценность действия в МППР зависит от текущего состояния

$$Q(s, a) = E[R_t | s_t = s, a_t = a] = E[\sum_{i=1}^{+\infty} r_{t+i} \gamma^{i-1} | s_t = s, a_t = a]$$

Общая задача

Уравнение Беллмана

Рекуррентное соотношение для ценности состояния.

$$
\begin{aligned}
V_\pi(s) &= E[\sum_{i=1}^{+\infty} r_{t+i}\gamma^{i-1}|s_t = s] \\
&= E[r_{t+1} + \gamma \sum_{i=1}^{+\infty} r_{t+i+1}\gamma^{i-1}] \\
&= E[r_{t+1}] + E[\gamma * V_\pi(s_{t+1})]
\end{aligned}
$$

Общая идея: пусть у нас есть оценка для $V_\pi(s)$. Давайте ходить туда, где эта оценка наилучшая, а потом обновлять эту оценку

Давайте использовать в качестве оценки какую-то вариацию на тему экспоненциального скользящего среднего

$$\Delta V(s_t) = \alpha_t(r_{t+1} + \gamma * V(s_{t+1}) - V(s_t))$$

Если $\sum \alpha_t^2 < +\infty$, $\sum \alpha_t = +\infty$, то оценки сходятся

Общая задача

Многошаговый метод временных разностей

Давайте использовать более точную оценку для целевой функции $\sum r_{t+i+1}\gamma^i$:

$$R_t = \sum_{i=1}^{n} r_{t+i}\gamma^{i-1}$$

Идея: можно обновлять значения в прошлом

Общая задача

Метод SARSA

Игра выглядит следующим образом:

- Инициализируем стратегию $\pi_1(a|s)$ и состояние среды s_1
- На каждом шаге:
- агент выбирает действие a_t из $\pi_t(a|s)$ (например, жадно)
- среда генерирует r_{t+1}, s_{t+1}
- агент разыгрывает еще один шаг a' из $\pi_t(a|s_{t+1})$
- обновляем $\Delta Q(s_t, a_t) = \alpha_t(r_{t+1} + \gamma Q(s_t, a') - Q(s_t, a_t))$

Литература и ссылки

- Sutton, Richard S.; Andrew G. Barto Reinforcement Learning: An Introduction. — MIT Press. 1998, 1998.
- Лекции Воронцова: http://tinyurl.com/zypufmc