

The K-Armed Bandit Problem

Summary:

Lesson 1: The k-Armed Bandit Problem

- The k-armed bandit

In the k-armed bandit problem, we have an agent who chooses between "k" actions and receives a reward based on the action it chooses.

$k=3$ Bandits actions

Rewards

- Action - Values

→ the values is the expected reward

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a] \quad \forall a \in \{1, 2, \dots, k\}$$
$$= \sum_r p(r|a) \cdot r$$

→ The goal is to maximize the expected reward

$$\arg \max_a q_*(a)$$

1. A k-armed Bandit Problem

- A k-armed Bandit được mô tả như sau:
 - Chúng ta có k hành động có xác suất thưởng là: $\{\theta_1, \theta_2, \dots, \theta_k\}$

- Tại mỗi thời điểm t , ta thực hiện một hành động và nhận được phần thưởng r .
- A_t : là tập các hành động, mỗi hành động đề cập đến sự tương tác với m .
- R : là một hàm phần thưởng. Trong trường hợp của k -armed Bandit, ta quan sát lấy phần thưởng r theo kiểu ngẫu nhiên.
- Ta có k hành động có các phần thưởng tương ứng. Sau mỗi lần lựa chọn, bạn sẽ nhận được phần thưởng từ phân phối xác suất cố định phụ thuộc vào hành động đã chọn.
- Mỗi hành động trong số k hành động có một phần thưởng mong đợi hoặc trung bình nếu hành động đó được chọn; thì được gọi là giá trị của hành động đó.
- Giá trị của một hành động a , ký hiệu $q_*(a)$ là phần thưởng kỳ vọng nếu a được chọn:

$$q_*(a) = \mathbb{E}[R_t | A_t = a]$$

A_t : là hành động được chọn ở thời điểm t

R_t : phần thưởng tương ứng

- Nếu biết giá trị của từng hành động, thì ta có thể luôn chọn hành động có giá trị cao nhất. Ta cho rằng không biết chắc chắn giá trị các hành động, và có thể ước tính. $Q_t(a)$: là giá trị ước tính của hành động a tại thời điểm t .

$$Q_t(a) \approx q_*(a)$$

- Exploitation (khai thác) and Exploration (khám phá)
 - Exploiting: lựa chọn một trong greedy actions. Khai thác sẽ tối đa hoá phần thưởng mong đợi trên một bước.
 - Exploring: lựa chọn một trong nongreedy actions. Khám phá có thể tạo ra tổng phần thưởng lớn hơn trong một thời gian dài.
 - Trong quá trình khám phá, phần thưởng sẽ thấp hơn trong thời gian ngắn nhưng cao hơn trong thời gian dài vì sau khi khám phá ra những hành động tốt hơn, ta có thể khai thác chúng nhiều lần.

2. Action – Value methods

- Các phương pháp ước tính giá trị của hành động và sử dụng ước tính để đưa ra quyết định lựa chọn hành động được gọi chung là phương pháp giá trị hành động (action – value methods)
- Một cách tự nhiên để ước tính là tính trung bình các phần thưởng thực sự nhận được:

$$Q_t(a) = \frac{\text{sum of reward when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i * \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$\mathbb{1}_{\text{predicate}}$: biểu thị biến ngẫu nhiên

$$\mathbb{I}_{predicate} = \begin{cases} 1 & \text{if predicate is true} \\ 0 & \text{if it is not} \end{cases}$$

- Quy tắc lựa chọn hành động đơn giản nhất: lựa chọn greedy actions. Nếu có nhiều hơn một hành động tham lam, thì việc lựa chọn sẽ được thực hiện theo một cách tùy ý nào đó, có thể là ngẫu nhiên.
- Greedy action selection method:

$$A_t \doteq \operatorname{argmax}_a Q_t(a)$$

3. Incremental Implementation

- Đặt Q_n biểu thị ước tính giá trị hành động sau khi được chọn $n-1$ lần, ta có :

$$Q_n \doteq \frac{R_1 + R_2 + \dots + R_{n-1}}{n-1}$$

R_i : biểu thị phần thưởng nhận được sau lần lựa chọn thứ i của hành động

- Khi cho trước Q_n và phần thưởng thứ n , giá trị trung bình của tất cả n phần thưởng sẽ được tính bằng:

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i = Q_n + \frac{1}{n} [R_n - Q_n]$$

$$NewEstimate \leftarrow OldEstimate + StepSize [Target - OldEstimate]$$

$[Target - OldEstimate]$: là lỗi trong ước tính

4. Tracking a Nonstationary Problem

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^n R_i = Q_n + \frac{1}{n} [R_n - Q_n]$$

- Đối với biểu thức trên, ở mỗi bước, xác suất phần thưởng không thay đổi theo thời gian.
- Tuy nhiên, sẽ hợp lý hơn nếu ta coi trọng phần thưởng gần đây hơn là phần thưởng trong quá khứ.
- Một trong những cách phổ biến nhất để thực hiện việc này là sử dụng step-size với kích thước không đổi:

$$Q_{n+1} \doteq Q_n + \alpha [R_n - Q_n] = (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i$$

- Đôi khi, để thuận tiện có thể thay đổi step-size từ bước này sang bước khác. Đặt $\alpha_n(a)$ biểu thị tham số step-size được sử dụng để đánh giá mức độ phần thưởng nhận được sau lựa chọn thứ n của hành động a .
- Để đảm bảo sự hội tụ, ta có điều kiện:

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \text{ and } \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

5. Upper – Confidence – Bound Action Selection

- UCB là một phương pháp có thể lựa chọn hành động để cân bằng giữa thăm dò và khai thác.

$$A_t \doteq \operatorname{argmax}_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

$c > 0$: tham số giúp kiểm soát mức độ thăm dò

$Q_t(a)$: đại diện cho phần thăm dò

$\sqrt{\frac{\ln t}{N_t(a)}}$: đại diện cho phần khai thác

t : số bước thời gian thực hiện

a : số bước thực hiện hành động a

→ khi hành động a được thực hiện càng nhiều lần, thì ta sẽ giảm thiểu phần thưởng của hành động a để có thể khám phá được nhiều hành động mới hơn.