Algoritmo UCB

Nelson Steven Sanabio Maldonado

Junio 2018

1 Algoritmo

La mecánica del algoritmo de confianza superior (UCB) es simple. En cada ronda, simplemente tiramos del lanzamiento que tiene la estimación de recompensa empírica más alta hasta ese punto más un término que es inversamente proporcional al número de veces que se ha jugado el lanzamiento. Más formalmente, defina $n_{i,t}$ como el número de veces que se ha jugado el lanzamiento i hasta el momento t. Defina $rt \in [0,1]$ ser la recompensa que observamos en el momento t. Define $I_t \in \{1...N\}$ para ser la elección del lanzamiento en el tiempo t. Entonces la estimación de recompensa empírica del lanzamiento i en el tiempo t es:

$$\mu_{i,t} = \frac{\sum_{s=0:\ I_s=i}^{t} r_s}{n_{i,t}} \tag{1}$$

UCB asigna el siguiente valor a cada lanzamiento i en cada momento t:

$$UCB_{i,t} := \mu_{i,t} + \sqrt{\frac{ln\, t}{n_{i,t}}}$$

El algoritmo UCB se da a continuación:

UCB

Input: N brazos, número de rondas $T \geq N$

1. Para $t = 1...N$, jugar lanzamiento t
2. Para $t = N+1...T$, juego de lanzamiento

$$I_t = arg_{i \in \{1...N\}} max\, UCB_{i,t-1}$$

Tenga en cuenta que estamos asumiendo (al menos en esta formulación) que jugaremos al menos N veces. Además, estamos actualizando implícitamente nuestra estimación empírica (1) cada vez que jugamos un lanzamiento. Observe que en el tiempo t, el algoritmo utiliza el UCBi, $t-1$, que se puede calcular utilizando observaciones realizadas hasta el tiempo $t-1$.

En un nivel intuitivo, el término adicional $\sqrt{\frac{ln\, t}{n_{i,t}}}$ nos ayuda a evitar siempre jugar el mismo brazo sin examinar otras armas. Esto es porque a medida que ni, t aumenta, $UCB_{i,t}$ disminuye. Tome el ejemplo de 2 lanzamiento: lanzamiento 1 con una recompensa fija de 0,25 y el lanzamiento 2 con una recompensa de 0-1 siguiendo una distribución de Bernoulli $\pi = 0,75$. Recuerde que la estrategia codiciosa (es decir, seleccionando $argmax_{i\in\{1\ldots N\}}\, \mu_{i,t}$) incurre en arrepentimiento lineal $R(T) = O(T)$ con probabilidad constante: con probabilidad 0.25, el brazo 2 produce recompensa 0, a que siempre seleccionaremos el lanzamiento 1 y nunca volveremos a visitar el lanzamiento 2. Si hacemos un seguimiento de UCB en esta situación, vemos que no tenemos este problema.

- (t = 1) El lanzamiento 1 se reproduce: $\mu_{1,1} = 0.25$.
- (t = 2) Se reproduce el lanzamiento 2: $\mu_{2,2} = 0$ (con probabilidad de 0,25 esto ocurre).
- (t = 3) Se reproduce el lanzamiento 1, porque $UCB_{1,2} = 0.25+\sqrt{ln\, 2} > UCB_{2,2} = 0 + \sqrt{ln\, 2}$
- (t = 4) Se reproduce el lanzamiento 2, porque $UCB_{1,3} = 0.25 + \sqrt{\frac{ln\, 3}{2}} \approx 0.9912 < UCB_{2,3} = 0 + \sqrt{ln\, 3} \approx 1.481$

1.1 Análisis de arrepentimiento dependiente de la instancia

Pero hay una razón más fundamental para la elección del término $\sqrt{\frac{ln\, t}{n_{i,t}}}$. Es un límite superior de alta confianza en el error empírico de $\mu_{i,t}$. Específicamente, para cada lanzamiento i en el tiempo t, debemos tener.

$$|\mu_{i,t} - \mu_i| < \sqrt{\frac{ln\, t}{n_{i,t}}} \tag{2}$$

con probabilidad de al menos $1 - \frac{2}{t^2}$. Hay dos límites útiles que podemos tomar inmediatamente de (2):

1. Un límite inferior para $UCB_{i,t}$. Con probabilidad al menos $1 - \frac{2}{t^2}$

$$UCB_{i,t} > \mu_i \tag{3}$$

2. Un límite superior para $\mu_{i,t}$ con muchas muestras. Dado que $n_{i,t} \geq \frac{4ln\, t}{\triangle_i^2}$, con probabilidad de al menos $1 - \frac{2}{t^2}$,

$$\mu_{i,t} < \mu_i + \frac{\triangle_i}{2} \tag{4}$$

(3) afirma que el valor UCB es probablemente tan grande como la verdadera recompensa: en este sentido, el algoritmo UCB es optimista. (4) declara que si se le dan suficientes (específicamente, al menos $\frac{4ln\, t}{\triangle_i^2}$) muestras, la estimación de la recompensa probablemente no exceda la recompensa verdadera en más de $\frac{\triangle_i}{2}$. Estos límites se pueden usar para mostrar que UCB rápidamente descubre un brazo subóptimo: