Algoritmo UCB

Nelson Steven Sanabio Maldonado

Junio 2018

1 UCB

1.1 Algoritmo

La mecánica del algoritmo de confianza superior (UCB) es simple. En cada ronda, simplemente tiramos del brazo 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 brazo. Más formalmente, defina $n_{i,t}$ como el número de veces que se ha jugado el brazo i hasta el momento t. Defina $r_t \in [0,1]$ ser la recompensa que observamos en el momento t. Define $I_t \in \{1...N\}$ para ser la elección del brazo en el tiempo t. Entonces la estimación de recompensa empírica del brazo 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 brazo 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 brazo t
2. Para $t = N+1...T$, juego de brazo

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

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 brazo. Observe que en el tiempo t, el algoritmo utiliza el $UCB_{i,t-1}$, 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 $n_{i,t}$ aumenta, $UCB_{i,t}$ disminuye. Tome el ejemplo de 2 brazos: el brazo 1 con una recompensa fija de 0,25 y el brazo 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 brazo 1 y nunca volveremos a visitar el brazo 2. Si hacemos un seguimiento de UCB en esta situación, vemos que no tenemos este problema.

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

1.2 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 brazo 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:

Lema 1.1. En cualquier punto t, si un brazo subóptimo i (es decir, $\mu_i < \mu^*$) se ha jugado para $n_{i,t} > \frac{4ln\, t}{\triangle_i^2}$ veces, entonces $UCB_{i,t} < UCB_{I^*,t}$ con probabilidad de al menos $1 - \frac{4}{t^2}$. Por lo tanto, para cualquier t,

$$P\left(I_{t+1} = i | n_{i,t} \geq \frac{4ln\, t}{\triangle_i^2}\right) \leq \frac{4}{t^2}$$

Lema 1.2. Deje que $n_{i,T}$ sea la cantidad de veces que el brazo i es tirado por el algoritmo de UCB ejecutado en la instancia $\Theta = \{\nu_1, \mu_1, ..., \nu_N, \mu_N\}$ del estocástico IID multi-armado bandido prbolem. Entonces, para cualquier brazo i con $\mu_i < \mu^*$,

$$\mathbb{E}[n_{i,T}] \leq \frac{4ln\, T}{\triangle_i} + 8$$

Teorema 1.3. Deje que $R(T, \Theta)$ denote el arrepentimiento del algoritmo de UCB en el tiempo T, por ejemplo $\Theta = \{\nu_1, \mu_1, ..., \nu_N, \mu_N\}$ del estocástico IID multi-armado bandido prbolem. Para todos los casos Θ, y todos $T \geq N$, el arrepentimiento esperado del algoritmo UCB está limitado como:

$$\mathbb{E}[R(T, \Theta)] \leq \sum_{i:\mu_i \leq \mu^*} \frac{4ln\, T}{\triangle_i} + 8$$

donde $\triangle_i = \mu^* - \mu_i$

1.3 Análisis de arrepentimiento independiente de la instancia

El teorema 1.3 da un límite superior en $\mathbb{E}[R(T, \Theta)]$ que es logarítmico en T. Esto está en una forma óptima: recuerdo de la última conferencia que cualquier algoritmo razonable debe sufrir en T esperado lamento total, no importa qué instancia Θ está dado.

Sin embargo, tenga en cuenta que el teorema 1.3 depende de una instancia específica de brazos, parametrizada por $\triangle_1 ... \triangle_N$. Dichos límites se denominan "dependientes de la instancia" o "límites dependientes del problema". Este límite implica directamente una muy buena pelea en el peor de los casos: por ejemplo, con $\triangle_i = lnT/T$, entonces el límite es lineal en T, que es tan malo como el algoritmo ϵ-greedy.

Pero se puede aplicar un simple truco al Teorema 1.3 para obtener el siguiente arrepentimiento "independiente de la instancia" (también conocido como "problema independiente" o "worst-case").

Teorema 1.4. Para todo $T \geq N$, el arrepentimiento total esperado logrado por el algoritmo UCB en la ronda T es

$$\mathbb{E}[R(T)] = 5\sqrt{NTlnT} + 8N$$

Proof. Solo para fines de análisis, divida los lanzamientos en dos grupos:

1. El grupo 1 contiene brazos "casi óptimos" con $\triangle_i < \sqrt{\frac{N}{T} ln\, T}$
2. El grupo 2 contiene brazos con $\triangle_i \geq \sqrt{\frac{N}{T} lnT}$

El arrepentimiento total es la suma de la pena de cada grupo. El remordimiento máximo total incurrido debido a los brazos de tracción en el Grupo 1 está limitado por

$$\sum_{i \in Grupo1} n_{i,T} \triangle_i \leq \left(\sqrt{\frac{N}{T} lnT} \right) \sum_{i \in Grupo1} n_{i,T} \leq T \sqrt{\frac{N}{T} lnT} = \sqrt{NTlnT}$$

donde se usó ese $\triangle_i \leq \frac{N}{T} lnT$ para todo i en el grupo 1, y el trivial límite $\sum_i n_{i,T} \leq T$ en el número total de tirones. A continuación, aplicamos el Lema 1.2 en cada brazo del Grupo 2 para unir el pesar esperado por

$$\sum_{i \in Grupo2} \mathbb{E}[n_{i,T}] \triangle \leq \sum_{i \in Grupo2} \frac{4ln\, T}{\triangle_i} + 8\triangle_i \leq \sum_{i \in Grupo2} 4\sqrt{\frac{Tln\, T}{N}} + 8 \leq 4\sqrt{NTln\, T} + 8N$$

donde en la primera desigualdad usamos eso para todo $i \in$ Grupo 2, $\sqrt{\frac{N}{T} ln\, T} \leq \triangle_i \leq 1$. Sumar las dos desigualdades da el resultado deseado.

References

[1] Peter Auer, Nicolo Cesa-Bianchi y Paul Fischer. Análisis en tiempo definido del problema de bandido múltiple , (2002), pp. 124-129.

[2] Karl Stratos, C.C. (2016).IEOR 8100-001, UCB Algorithm, Worst-Case Regret Bound. Learning and Optimization for Sequential Decision Making