# 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:

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

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 $t = N + 1...T, juegodelanzamiento$

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

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 $t-1$.