Rapport sur l'implémentation de l'algorithme UCB1

11 juin 2025

1 Principe de l'algorithme

L'algorithme UCB1 (*Upper Confidence Bound*) est une solution classique au problème du bandit manchot (*multi-armed bandit problem*), qui pose un dilemme fondamental entre exploration et exploitation en apprentissage automatique.

Intuitivement, imaginons que nous sommes dans un casino avec un nombre fini K de machines à sous (bras). Chaque machine donne une récompense (positive ou négative) de manière aléatoire selon une distribution de probabilité inconnue. L'objectif est de maximiser le gain cumulé en choisissant intelligemment entre :

— **Explorer** : tester des machines peu connues pour estimer leur potentiel ;
— **Exploiter** : utiliser les machines que l'on pense être les meilleures.

L'algorithme UCB1 propose un compromis efficace entre ces deux objectifs, en tenant compte à la fois de la moyenne empirique des récompenses et de l'incertitude associée à chaque bras.

2 Modélisation mathématique du problème

Soit K le nombre de bras (actions possibles). À chaque instant t, l'agent sélectionne un bras et observe une récompense.

— Pour chaque bras $i = 1, \ldots, K$, on observe $x_{i,t}$, une réalisation d'une variable aléatoire $X_{i,t} \in [0, 1]$, issue d'une distribution inconnue ν_i avec espérance μ_i.
— On définit le regret du choix du bras i après n tours comme :

$$R(n) = n\mu^* - \mu_i \sum_{t=1}^{n} \mathbb{E}[T_t(n)]$$

où $\mu^* = \max_{i=1,\dots,K} \mu_i$ est la récompense moyenne maximale et $T_i(n)$ est le nombre de fois que le bras i a éte choisit parmis les n tours.

— L'objectif est de minimiser ce regret en choisissant, à chaque étape, le bras avec le meilleur compromis entre moyenne estimée et incertitude ie avec un regret relativement petit.

Dérivation de la borne de confiance dans UCB1

On part de l'inégalité de Hoeffding pour des variables dans $[0, 1]$:

$$\Pr\big(\hat{\mu}_i(t) + \varepsilon < \mu_i\big) \;=\; \Pr\big(\hat{\mu}_i(t) - \mu_i \le -\varepsilon\big) \;\le\; \exp\big(-2\, N_i(t)\, \varepsilon^2\big).$$

Pour garantir que cette probabilité soit au plus t^{-4}, on fixe

$$\exp\big(-2\, N_i(t)\, \varepsilon^2\big) \;=\; t^{-4}.$$

En prenant les logarithmes, on obtient

$$-2\, N_i(t)\, \varepsilon^2 = -4\, \ln t \quad\Longrightarrow\quad \varepsilon^2 = \frac{2\, \ln t}{N_i(t)} \quad\Longrightarrow\quad \varepsilon = \sqrt{\frac{2\, \ln t}{N_i(t)}}.$$

Ainsi, la borne supérieure de confiance s'écrit

$$\mathrm{UCB}_i(t) \;=\; \hat{\mu}_i(t) + \sqrt{\frac{2\, \ln t}{N_i(t)}}.$$

Le principe même de l'algorithme repose sur le fait de choisir a chaque etapes le bras qui maximise cette borne.

3 Algorithme UCB1

L'algorithme UCB1 fonctionne de la manière suivante :

1. Initialiser : jouer une fois chaque bras $i \in \{1, \dots, K\}$.
2. Pour chaque étape $t > K$, choisir le bras :

$$I_t = \arg \max_{i \in \{1,\dots,K\}} \left(\hat{\mu}_i(t) + \sqrt{\frac{2 \ln t}{N_i(t)}} \right)$$

où :

— $\hat{\mu}_i(t)$ est la moyenne empirique des récompenses du bras i jusqu'à l'instant t,
— $N_i(t)$ est le nombre de fois que le bras i a été joué jusqu'à t.

3. Mettre à jour les statistiques du bras sélectionné.

L'idée est que le deuxième terme (*bonus de confiance*) encourage l'exploration des bras peu essayés, tandis que la moyenne empirique favorise l'exploitation des bras prometteurs.

4 Preuve de la borne du regret pour UCB1

On cherche à établir une borne supérieure sur le **regret cumulatif** de l'algorithme UCB1. Le regret est défini comme :

$$R(n) = \sum_{i:\mu_i<\mu^*} \Delta_i \cdot \mathbb{E}[T_i(n)]$$

où :

— μ_i est l'espérance du bras i,
— $\mu^* = \max_j \mu_j$ est l'espérance optimale,
— $\Delta_i = \mu^* - \mu_i$ est l'écart de performance,
— $T_i(n)$ est le nombre de fois que le bras i a été sélectionné jusqu'à l'instant n.

L'objectif est de montrer que pour tout bras sous-optimal i, on a :

$$\mathbb{E}[T_i(n)] \leq \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

Idée de la démonstration

L'algorithme sélectionne à chaque étape t le bras i maximisant :

$$\hat{\mu}_i(t) + \sqrt{\frac{2 \ln t}{T_i(t)}}$$

Un bras sous-optimal est sélectionné seulement si l'un des événements suivants se produit :

1. La moyenne empirique du bras optimal est significativement sous-estimée ;
2. La moyenne empirique du bras i est surestimée ;
3. Le terme de confiance fausse l'estimation en faveur de i.

En appliquant l'inégalité de Hoeffding, on montre que ces événements sont rares, et qu'après un certain nombre de tirages (environ $\frac{8 \ln n}{\Delta_i^2}$), le bras i ne sera plus choisi sauf avec faible probabilité.

Ainsi, le regret cumulé est borné par :

$$R(n) \leq \sum_{i:\mu_i<\mu^*} \left(\frac{8 \ln n}{\Delta_i} + \Delta_i \left(1 + \frac{\pi^2}{3} \right) \right)$$

Ce qui garantit que le regret de UCB1 croît de manière logarithmique, ce qui est optimal pour des distributions stationnaires et indépendantes.