ENSAE

INSTITUT POLYTECHNIQUE DE PARIS

Algorithmes Bandits

Flore Sentenac

Doctorante au CREST, sous la direction de Vianney Perchet

Première formulation du problème

Thompson, 1933, essai clinique, pilule jaune vs. pilule rouge

Guérison

Sans effet

Quelle pilule pour les patients suivants ?

Origine du nom

Les bandits manchots de Las Vegas

Une option $\rightarrow$ un bras

Choisir une option $\rightarrow$ tirer un bras

Explorer ou Exploiter ?

Une pilule est efficace à 40%, l'autre à 60%. On ne sait pas quelle pilule est efficace à 60%.
On veut traiter 10000 patients.

EXPLOITATION

Pilule jaune pour tous les prochains patients.

- Probabilité de 30% que la jaune soit la mauvaise
- 5380 patients guéris en espérance

EXPLORATION

Pilule jaune pour la moitié des patients, pilule rouge pour l'autre.

- La moitié des patients prennent la mauvaise pilule
- 5000 patients guéris en espérance

Bonne pilule à tout le monde: 6000 patients guéris en espérance

Explorer *puis* Exploiter

- EXPLORATION: 200 personnes prennent la pilule jaune, 200 personnes la pilule rouge. Calcul des moyennes empiriques de succès: $\hat{\mu}_{\text{rouge}}$ et $\hat{\mu}_{\text{jaune}}$.
- EXPLOITATION: Si $\hat{\mu}_{\text{rouge}} < \hat{\mu}_{\text{jaune}}$, donner la pilule jaune au 9600 patients restants.

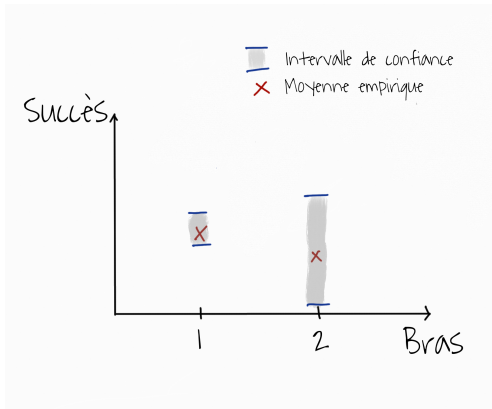
Si $\mu_{\text{rouge}} > \mu_{\text{jaune}}$, alors $\mathbb{P}(\hat{\mu}_{\text{rouge}} < \hat{\mu}_{\text{jaune}}) \leq 0.1$: on se trompe rarement.

En espérance, plus de 5768 patients guéris.

Explorer en exploitant: Algorithme UCB

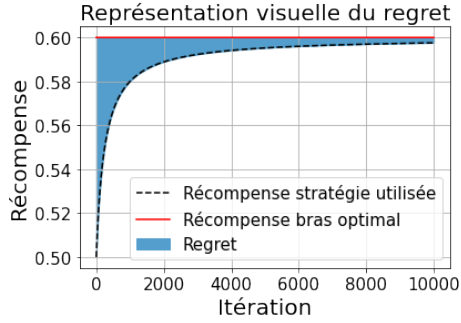
Principe

1. Calculer les moyennes empiriques de succès de chaque bras
2. Construire des intervalles de confiance
3. Tirer le bras avec la plus grande borne supérieure



P. Auer et al., 2002

Regret



- K options = K bras
- Récompense moyenne du bras $k \in [K]$: μ_k
- Ces valeurs sont inconnues pour le joueur
- A l'itération t, le joueur choisit le bras a_t
- Reçoit la récompense $\mu_{a_t} + \epsilon_t$, ϵ_t un bruit blanc

OBJECTIF

Maximiser la récompense espérée obtenue en T itérations, ou de manière équivalente, minimiser le regret:

$$R(T) := T \max_{k \in [K]} \mu_k - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{a_t}\right]$$

Au delà du bandit à K bras

- Tirer plusieurs bras à la fois $\rightarrow$ bandits combinatoires
- Jouer contre des concurrents $\rightarrow$ bandits à plusieurs joueurs
- Utilisation des méthodes bandits pour des algorithmes d'allocation séquentiels

Merci pour votre attention !