Algorithme et modèles de bandits

06/04/2018

Benjamin Phan Charlotte Pasquier

Origine

- Cet algorithme apparu dans les années 1930 était initialement utilisé en medecine afin d'identifier le meilleur traitement parmi K médicaments.
- X_t=1 si le patient est guéri, 0 sinon.
- A chaque tour t, un traitement (nommé bras) est testé.
- Le meilleur traitement est celui qui a le meilleur taux de guérison, à horizon T il s'agit de celui qui minimise le reget :

$$R^{\pi}(T) = \mu^* T - E\left[\sum_{t=1}^{T} X_t\right]$$

Quelle utilisation aujourd'hui?

- Aujourd'hui l'algorithme est principalement utilisé dans le cadre du marketing quantitatif : algorithmes de recommandation, publicités en ligne etc...
- Compromis :
  -exploration
  -exploitation
- Plusieurs objectifs peuvent être envisageables :
  -minimisation du regret
  -identification du meilleur bras à budget ou intervalle de confiance fixé

Méthodologie

- Que l'on soit dans le cadre de la minimisation du regret ou de l'identification du meilleur bras, l'algorithme doit à chaque étape décider quel sera le bras choisi.

- Les algorithmes UCB selectionne le bras qui obtient la plus grande borne supérieure de l'intervalle de confiance autour de μ_a.

Borne (1/2)

Objectif : obtenir le plus petit regret possible.

- Lai et Robbins [1985] montrent qu'il est possible d'obtenir une borne **logarithmique** sur le regret.
- Toute stratégie π uniformement efficiente, c'est à dire dont le regret est petit pour tous les modèles de bandits de la classe, vérifie $\forall \mu \in I^k$

$$\lim_{T \to \infty} \inf \frac{R_\mu^\pi(T)}{log(T)} \geq \sum_{a:\mu_a < \mu^*} \frac{1}{d(\mu_a, \mu^*)}$$

- Un algorithme est dit **asymptotiquement optimal** si $\forall \mu \in I^k$ son regret est limité par :

$$log(T) \sum_{a:\mu_a < \mu^*} \frac{1}{d(\mu_a, \mu^*)}$$

Borne (2/2)

- A horizon fini et pour certains μ la borne présentée n'est pas très intéréssante, dans ce cas on préfère se référer à la borne de Cesa Bianchi [2012] dont le reget est en $\sqrt{KT}$.

$$\inf_{\pi} \sup_{\mu \in I} R_{\mu}^{\pi}(T) \geq \frac{1}{20}\sqrt{KT}$$

Quels algorithmes ?

UCB1 :

- Pour T fini, l'algorithme UCB1 permet d'obtenir une borne sur le regret [Auer et al., 2002]. Le bras joué à chaque tour est :

$$A_{t+1} = argmax_{a \in \{1,\dots,K\}} \hat{\mu}_a(t) + \frac{2\sigma^2 log(t)}{N_a(t)}$$

- Dans le cadre d'une distribution de type Bernouilli, le regret obtenu est alors de l'ordre :

$$\sum_{a:\mu_a<\mu^*}^{K} \frac{1}{2(\mu^*-\mu_a)} log(T) + O(\sqrt{log(T)})$$

kl-UCB :

- Pour une famille exponentielle à un paramètre, l'algorithme kl-UCB est **asymptotiquement optimal** [Cappé et al., 2013] et le bras joué à chaque tour est :

$$A_{t+1} = argmax_{a \in \{1,\dots,K\}} max\{q : N_a(t) d(\hat{\mu}_a, q) \leq log(t) + 3log(log(T))\}$$

Choix du bras

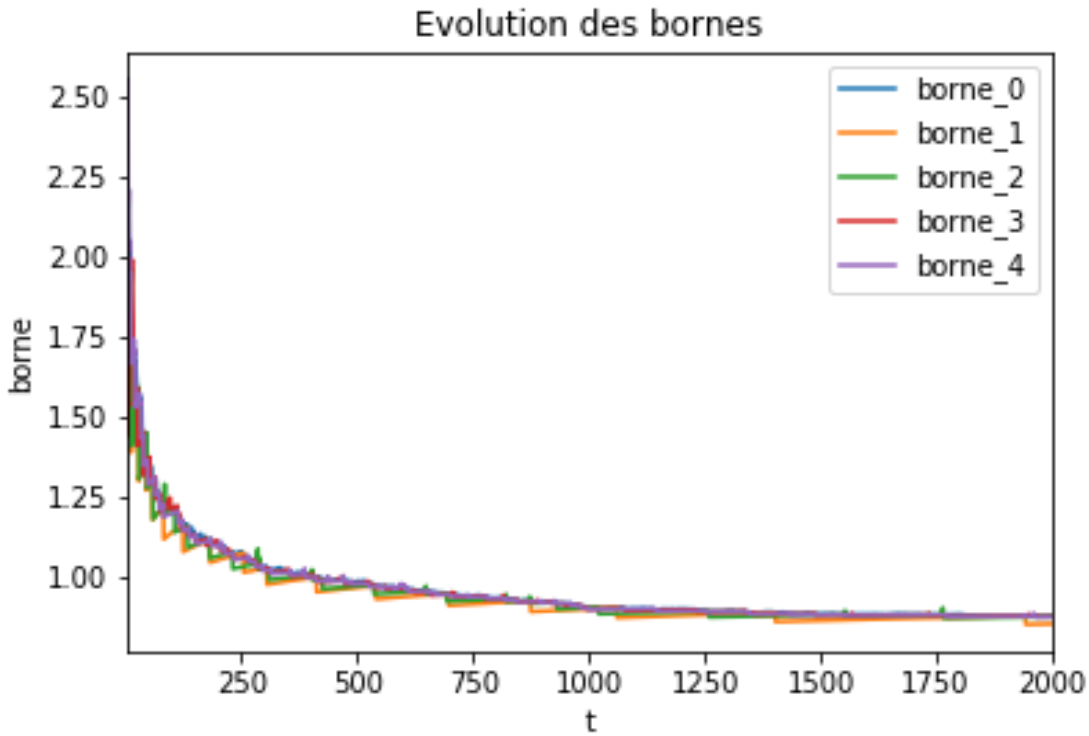

Figure: L'algorithme UCB choisit le bras avec la borne supérieure de l'intervalle de confiance la plus élevée

Regret

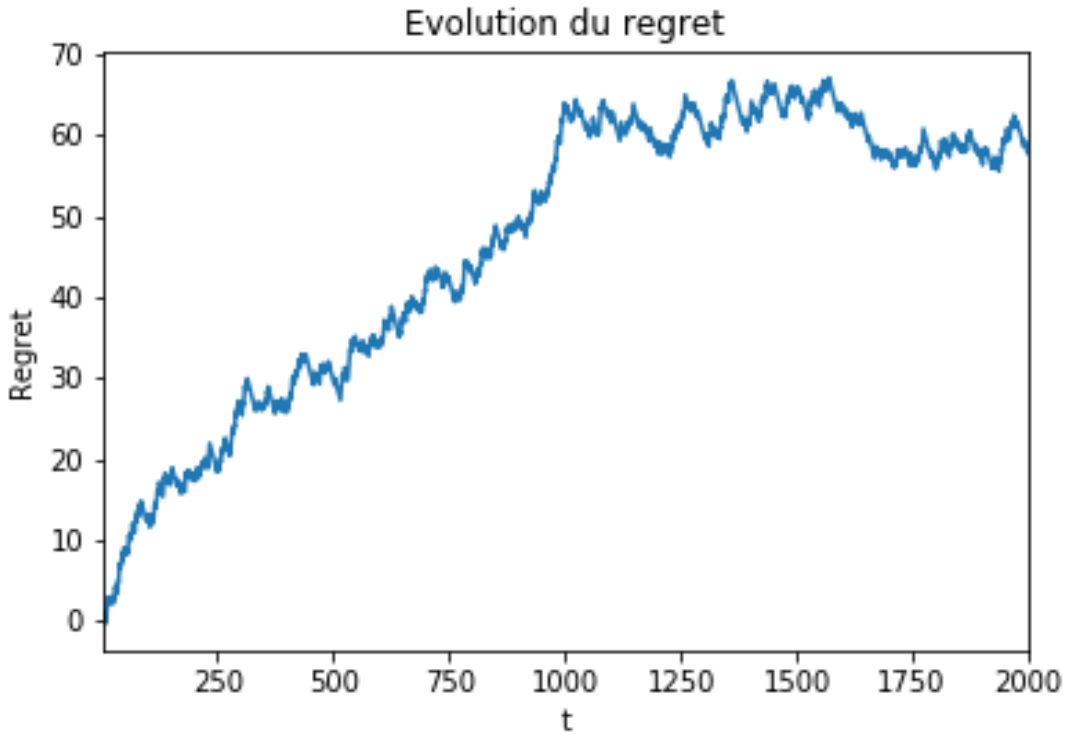

Figure: Le regret croît de moins en moins vite puis quasiment plus : le meilleur bras est alors choisit quasiment tout le temps

L'alternative Bayesienne

L'article présente également succinctement deux autres types d'algorithmes **asymptotiquement optimaux** (dans le cadre de modèle à 1 paramètre) issus de la littérature bayesienne.

Bayes - UCB :

- A chaque tour le bras qui a le quantile d'ordre $1 - \frac{1}{t}$ de la distribution posterieure de μ_a le plus élevé est sélectionné. [Kaufmann et al., 2012a]

Thompson Sampling :

- A chaque tour un bras est selectionné selon sa probabilité d'être optimal.[Thompson, 1933]

Cadre général

Objectif : trouver le meilleur bras avec grande probabilité.

- L'algorithme des bandits dans le cadre de l'identification du meilleur bras premet d'aboutir à :
  - une règle de décision : elle détermine quel sera le bras A_t choisi à chaque tour
  - une règle d'arrêt τ : elle détermine le temps nécéssaire à l'identification du meilleur bras
  - une règle de recommandation $\hat{a}_t$: elle spécifie le meilleur bras
- Soit S l'ensemble contenant tous les modèles de bandits ayant un seul bras optimal, uns stratégie est dite δ-PAC si :

$$\forall \mu \in S, P_\mu \left(\hat{a}_\tau = a^*(\mu)\right) \geq 1 - \delta$$

Bornes (1/2)

- Soit $\mu \in S$ avec S une famille exponentielle, $Alt(\mu) = \{\lambda \in S : a^*(\lambda) \neq a^*(\mu)\}$, l'ensemble contenant tous les modèles alternatifs tels que le meilleur bras est différent de celui choisi sous μ et w_a le poids associé à chaque bras [Garivier and Kaufmann, 2016].

  Alors tout algorithme δ-PAC satisfait :

$$\mathbb{E}_\mu[\tau] \geq T^*(\mu) kl(\delta, 1-\delta)$$

$$\text{avec } T^*(\mu)^{-1} = \sup_{w \in \Sigma_K} \inf_{\lambda \in Alt(\mu)} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

Bornes (2/2)

- Pour fonctionner l'algorithme doit trouver selon quelles proportions chacun des bras seront tirés. La proportion optimale est la suivante :

$$w^*(\mu) \in \arg\max_{w \in \Sigma_K} \inf_{\lambda \in Alt(\mu)} \sum_{a=1}^{K} w_a d(\mu_a, \lambda_a)$$

Les $w^*(\mu)_a$ peuvent être calculés explicitement par recherche dichotomique, ou, méthode de Newton

Quels algorithmes ?

Track-and-Stop :

- L'algorithme Track-and-Stop s'appuie sur deux règles :

  -la **règle de "suivi"** : elle s'assure que la proportion empirique de tirage de chacun des bras converge bien vers $w_a^*(\mu)$.

  -la **règle d'arrêt de Chernoff** : elle détermine à quel moment arrêter l'algorithme dès lors que l'on est suffisamment proche de la borne inférieure trouvée précedemment.

Règle de suivi (1/2)

- Le choix du bras tiré à chaque tour résulte d'un compromis entre :
  -exploration forcée : chaque bras doit être suffisamment exploré
  -cohérence avec les proportions optimales estimées par $w_a^*(\hat{\mu}(t))$
- Soit $F_t = \{a : N_a(t) < \sqrt{(t)} - K/2\}$. Le bras choisi au tour t+1 est:

$$A_{t+1} \in \begin{cases} \underset{a \in F_t}{\arg\min} & N_a(t) & \text{si } F_t \neq \emptyset \\ \underset{1 \leq a \leq K}{\arg\max} & w_a^*(\hat{\mu}(t)) - N_a(t)/t & \text{sinon} \end{cases}$$

Règle de suivi (2/2)

- La règle de suivi assure la convergence des proportions observées vers les poids optimaux à horizon infini :

$$\mathbb{P}_\mu \left(\lim_{t \to \infty} \frac{N_a(t)}{t} = w_a^*(\mu) \right) = 1$$

Règle d'arrêt de Chernoff

- La règle d'arrêt de Chernoff est la suivante :

$$\tau_\delta = inf\{t \in \mathbb{N} : \hat{Z}(t) > \beta(t,\delta)\}$$

avec $\hat{Z}(t) = \frac{\ell(X_1,\ldots,X_t;\hat{\mu}(t))}{\sup\limits_{\lambda \in Alt(\hat{\mu}(t))} \ell(X_1,\ldots,X_t;\lambda)}$

- Dans le cadre des lois de Bernouilli il est possible de montrer en utilisant les tests de rapport de vraisemblance généralisé (GLRT) que $\beta(t,\delta) = \frac{2(K-1)t}{\delta}$.
- Intuitivement, l'algorithme s'arrête dès lors que le GLRT rejette l'hypothèse $\mu_{\hat{a}(t)} < \mu_b$ pour tout bras b .

Simulations

On implémente les algorithmes UCB1, kl-UCB$^+$ et Track-and-Stop.

- Le regret obtenu par l'algorithme Track-and-Stop est supérieur à ceux obtenus avec les algorithmes UCB
- kl-UCB$^+$ a un regret plus faible que UCB1 en moyenne mais a une plus grande variance
- l'identification du bon bras est moins fréquente que prévue avec le Track-and-Stop

Conclusion

Deux objectifs principaux:
-minimisation du risque (algorithme de type UCB)
-identification du meilleur bras le plus rapidement possible (Track-and-Stop).

D'un point de vue computationnel un algorithme UCB peut s'avérer plus efficace que Track-and-Stop pour identifier le meilleur bras.