Algorithmes UCB et ε-Glouton

Enseignant : L. Ralaivola
Date : 8 novebmre 2017

1 Bandits à K bras et dilemme exploration/exploitation

1.1 Description du problème

Supposons que nous sommes dans le contexte des bandits à K bras, tels que décrits précédemment. L'objectif d'un joueur, dont on va supposer qu'il va faire T tirages de bras, est de trouver la meilleure séquence de bras à choisir pour maximiser ses gains. La difficulté de sa tâche provient de ce qu'il ne connaît pas le bras le plus profitable et que son objectif n'est pas simplement d'identifier ce bras, mais de le trouver en consacrant le moins de tirages possibles à des bras sous-optimaux, de sorte que son gain final soit le plus élevé possible. Cet objectif conduit l'agent au-devant du dilemme exploitation/exploration : au bout de t tirages de bras, il peut avoir une idée, en calculant les moyennes des gains rapportés par chaque bras, du bras le plus prometteur et choisir l'une des 2 options suivantes :

exploitation : ne plus jouer que sur le bras identifié comme le meilleur jusqu'à présent, avec l'hypothèse que c'est effectivement le meilleur de tous les bras ;

exploration : continuer à jouer d'autres bras, et éventuellement des bras non encore joués jusqu'à présent, pour se laisser la possibilité de découvrir des bras encore meilleurs que celui qui avait été identifié comme tel, au risque de faire des tirages de bras sous-optimaux.

Les algorithmes présentés dans ce document ont précisément été conçus pour définir des séquences de tirages de bras maximisant le gain au bout de T tirages : ils apportent donc des réponses au dilemme exploration/exploitation qui vient d'être décrit.

1.2 Formalisation : minimisation du regret

Pour décrire les algorithms ε-glouton et UCB qui sont au cœur de ce document, il est nécessaire de formaliser le problème qui nous intéresse et, en particulier, de montrer comment le dilemme exploration/exploitation s'écrit mathématiquement.
Pour rappel, une manière de décrire un bandit est comme suit :

- à chaque bras k correspond une variable aléatoire X_k de moyenne $\mu_k \doteq \mathbb{E}X_k$;
- les variables aléatoires $X_1, \ldots, X_K$ sont indépendantes (deux à deux) et prennent des valeurs dans $[0;1]$;
- la récompense $X_{I_t}^t$ obtenue lors du tirage du bras I_t au temps t est une copie indépendante de X_{I_t} (i.e., si $I_t = k$, alors $X_{I_t}^t$ suit la même loi que X_k et la récompense obtenue à l'instant t est une réalisation de cette variable aléatoire).

L'objectif d'un agent (ou joueur) $\mathcal{A}$ qui fait T tirages $I_1, I_2, \ldots, I_T$ sur ce bandit à K bras est de maximiser son gain

$$\mathcal{G}(\mathcal{A}) \doteq \sum_{t=1}^{T} X_{I_t}^t.$$

Un autre critère que le gain pour mesurer la stratégie de $\mathcal{A}$ est le regret

$$\widehat{\mathcal{R}}(\mathcal{A}) \doteq \max_{k=1,\ldots,K} \sum_{t=1}^{T} X_k^t - \mathcal{G}(\mathcal{A}) = \max_{k=1,\ldots,K} \sum_{t=1}^{T} X_k^t - \sum_{t=1}^{T} X_{I_t}^t,$$

qui mesure la différence entre les gains rapportés par le « meilleur bras » sur la séquence de tirages et le gain obtenu par $\mathcal{A}$. L'agent $\mathcal{A}$ doit *minimiser* son regret. Le regret est une quantité qui n'est pas facile à manipuler et deux autres quantités lui sont souvent préférées pour l'analyse de stratégies : le regret espéré $\mathbb{E}\widehat{\mathcal{R}}(\mathcal{A})$ et le *pseudo-regret* $\mathcal{R}(\mathcal{A})$,

$$\mathbb{E}\widehat{\mathcal{R}}(\mathcal{A})) \doteq \mathbb{E}\left[\max_{k=1,\ldots,K} \sum_{t=1}^{T} X_k^t - \sum_{t=1}^{T} X_{I_t}^t\right]$$
$$\mathcal{R}(\mathcal{A}) \doteq \max_{k=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{T} X_k^t - \sum_{t=1}^{T} X_{I_t}^t\right],$$

où il faut noter que l'espérance est prise par rapport aux variables aléatoires X_k^t et $X_{I_t}^t$ mais aussi par rapport aux variables aléatoires I_t, qui correspondent aux choix de l'agent – ces choix sont aléatoires car, sauf à utiliser des séquences de tirages prédéfinies, ils dépendent des récompenses obtenues, elles aussi aléatoires.
Les algorithmes présentés ci-dessous sont des algorithmes qui minimisent le pseudo-regret, qui peut se ré-écrire

$$\widehat{\mathcal{R}}(\mathcal{A}) = T\mu^* - \sum_{t=1}^{T} \mathbb{E}_{I_t} \mu_{I_t},$$

où $\mu^* \doteq \max_k \mu_k$, correspond à la récompense moyenne la plus élevée de tous les bras (plusieurs bras peuvent correspondre à cette récompense moyenne). Le pseudo-regret vient rendre compte de ce qu'on perd à ne pas jouer systématiquement un bras dont l'espérance est la plus élevée.
Les algorithmes présentés par la suite proposent des stratégies pour les tirages de bras qui visent à minimiser le pseudo-regret.

2 Algorithmes

Avant de présenter en détail les algorithmes ε-glouton et `UCB`, nous introduisons les notations suivantes :

— $T_k(t)$ désigne le nombre de fois que le bras k a été choisi par l'agent jusqu'au temps t ;

— $\widehat{\mu}_k^t$ désigne la moyenne empirique des récompenses obtenues sur le bras k au temps t :

$$\widehat{\mu}_k^t \doteq \frac{1}{T_k(t)} \sum_{\tau=1}^{t} \mathbb{I}(I_\tau = k) X_k^\tau.$$

Comme nous allons le voir, c'est sur la valeur de ces quantités (i.e. les réalisations de ces variables aléatoires) au temps $t-1$ que repose le choix I_t du bras au temps t.

2.1 Algorithme ε-glouton

L'algorithme ε-glouton est un algorithme très simple, décrit par exemple dans [6]. Dans sa version élémentaire, la stratégie déterminée par ε-glouton repose sur la donnée d'un réel ε tel que $0 < \varepsilon < 1$; cette stratégie suggère

— avec probabilité $1-\varepsilon$, de choisir le bras I_t dont la récompense au temps $t-1$ est la plus élevée, i.e. :

$$I_t \doteq \operatorname*{argmax}_k \widehat{\mu}_k^t = \operatorname*{argmax}_k \frac{1}{T_k(t)} \sum_{\tau=1}^{t} \mathbb{I}(I_\tau = k) X_k^\tau,$$

— avec probabilité ε, de choisir un bras I_t parmi les bras non-optimaux, avec une probabilité uniforme.

L'algorithme est initialisé en faisant 1 tirage de chaque bras sur les K premières itérations, la procédure ci-dessus n'étant activée qu'à partir du moment où chacun des bras a été tiré au moins une fois.

Une version améliorée de l'algorithme consiste à utiliser un paramètre ε_t adaptatif (en lieu et place de ε) prenant la forme $\varepsilon = K/(d^2 t)$: si d est judicieusement choisi (petit), alors il est possible de montrer que le regret est de l'ordre de $K \ln T/d^2$ (voir [3]).

On remarque que ε est un paramètre qui détermine la propension de l'agent à « explorer » : plus ε est grand, plus l'agent a tendance à choisir le prochain bras sans tenir compte des informations glanées dans le passé, inversement, plus ε est petit, plus l'agent fera reposer sa stratégie sur l'exploitation du bras qu'il aura jugé comme le plus prometteur au temps t.

2.2 Algorithme Upper Confidence Bound

L'algorithme *Upper Confidence Bound* ou UCB [1, 2] est construit à partir de l'inégalité de Chernoff/Hoeffding [4, 5], qui quantifie la probabilité avec laquelle la moyenne empirique de réalisations indépendantes d'une variable aléatoire s'éloigne de son espérance. Sans entrer dans les détails mathématiques qui donnent lieu au calcul de I_t, la stratégie UCB est paramétrée par un réel $\alpha > 0$ et suggère de calculer les indices $\mathrm{UCB}_k(t)$ pour tous les bras à chaque étape t suivant la formule

$$\mathrm{UCB}_k(t) \doteq \widehat{\mu}_k^{(t-1)} + \sqrt{\frac{\alpha \ln t}{2T_k(t-1)}}$$

et de choisir I_t selon :

$$I_t = \underset{k=1,\dots,K}{\operatorname{argmax}} \, \texttt{UCB}_k(t). \tag{1}$$

Comme précédemment, cette procédure n'a de sens que si chacun des K bras a été tiré au moins une fois ; l'initialisation de UCB repose donc sur le tirage aléatoire de chacun des K bras au cours des K premières itérations.
Ici, le degré d'exploration dépend du paramètre α :

- lorsque α est petit, la stratégie induite par UCB est une stratégie d'exploitation, puisqu'alors I_t correspond au bras dont la récompense moyenne empirique est la plus élevée ;
- inversement, lorsque α est grand, le second terme de l'équation (1) a une importance plus grande et UCB favorise le choix de I_t en fonction de ce second terme essentiellement ; les bras k ayant une valeur de $T_k(t-1)$ sont donc privilégiés (et il faut comprendre que la stratégie induite invite à glaner de l'information sur les bras qui n'ont pas suffisamment été choisis au temps t).

Il est possible de montrer que le pseudo-regret d'UCB est de l'ordre de $\ln T$.

3 Conclusion

Nous avons présenté le dilemme exploration/exploitation lié au problème des bandits. Après avoir présenté la notion de regret, et en particulier celle de pseudo-regret, nous avons décrit 2 algorithmes, ε-glouton et UCB, dont l'objectif est d'apporter une réponse à ce dilemme et qui donnent des procédures extrêmement simples pour le choix des séquences $I_1, \dots, I_T$ de bras à tirer.
La programmation de ces 2 algorithmes requiert essentiellement de trouver la bonne structure de données pour représenter un agent $\mathcal{A}$ qui, comme le préconisent les équations présentées ci-dessus, n'a besoin de mémoriser que quelques éléments sur les tirages passés pour prendre ses futures décisions (essentiellement : les quantités $T_k(t)$ et les moyennes des gains $\widehat{\mu}_k^t$).

Références

[1] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem,. *Machine Learning Journal*, 47(2-3) :235–256, 2002.

[2] P. Auer and R. Ortner. Ucb revisited : Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61 :55–65, 2010.

[3] S. Bubeck and N. Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, volume 5 of *Foundation and Trends in Machine Learning*. NOW, 2012.

[4] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematic Statistics*, 23 :493–507, 1952.

[5] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963.

[6] R. S. Sutton and A. G. Barto. *Reinforcement Learning : An Introduction*. MIT Press, 1998.