Master Recherche IAC
Option 6
Multi-Armed Bandits

Odalric-Ambryn Maillard
LRI

Jan 9th, 2015

Overview

Multi-Armed Bandits

Introduction

- Pourquoi
- Définitions
- Echec de l'algorithme glouton

Algorithmes

- Upper Confidence Bound (UCB)
- Kullback-Leibler UCB
- BESA

Pourquoi

Lai, Robbins 85

In a casino, one wants to maximize one's gains *while playing*.

Lifelong learning

Exploration vs **Exploitation** Dilemma

- Play the best arm so far ? **Exploitation**
- But there might exist better arms... **Exploration**

Vocabulaire

- K options, ou bras (arms)
- Ces options sont indépendantes
- L'option i retourne une récompense r tirée selon la distribution ν_i
  Par exemple, ν_i retourne 1 avec une certaine probabilité μ_i, et 0 sinon (variable de Bernoulli).

Objectifs possibles

- Identifier le meilleur bras (option i^* tel que $\mu_{i^*} = max_i \mu_i$.
- Trouver une stratégie (à chaque pas de temps t on choisit un bras i_t) maximisant la somme des récompenses.

Exemples d'application

- Choix des publicités affichées à l'internaute.
  $r = 1$ si l'internaute clique sur la pub et 0 sinon.

- Choix du traitement pour une maladie donnée.
  $r = 1$ si le malade est guéri et 0 sinon.

The multi-armed bandit (MAB) problem

Algorithme général

Unknown parameters: K unknown probability distributions on $[0, 1]$
Known parameters: the set of arms $1 \dots K$, the number of rounds T

For each round $t = 1, 2, \dots, T$

(1) the learner chooses $i_t \in 1 \dots K$ according to its own strategy.

(2) the learner incurs and observes the reward $r_t \sim \nu_{i_t}$ independently from the past given rewards.

T: horizon temporel

Dans le cas où T est inconnu, on parle d'algorithme *anytime*

Overview

Objectif

$$\textbf{Goal: Maximize } \sum_{u=1}^{t} r_u$$

$\Leftrightarrow$

$$\textbf{Minimize Regret } (t) = \sum_{u=1}^{t} (r \sim \nu^* - r_u)$$

La différence par rapport à ce que l'oracle (qui connait i^*) aurait gagné.

Intérêt

Le fait de se comparer à l'oracle est plus neutre: si le problème est difficile, l'oracle lui-même n'aurait pas fait très bien. Le regret mesure combien on dégrade par rapport à ce qu'on aurait pu faire au mieux.

Lemme

Notations

- $n_{i,t}$: nombre de tirages du bras i jusqu'au round t
- $\hat{\mu}_{i,t}$ est la moyenne empirique du bras i:

$$\hat{\mu}_{i,t} = \frac{1}{n_{i,t}} \sum_{u=1}^{t} r_u . \mathbb{I}_{i_u=i}$$

où $\mathbb{I}_e = 1$ qd e est vrai et 0 sinon

- $\mu_i = \mathbb{E}[\nu_i]$
- Δ_i: marge du bras $i = \mu^* - \mu_i$

Alors

$$\mathbb{E}[Regret_t] = \sum_{i=1}^{K} \Delta_i \mathbb{E}[n_{i,t}]$$

L'approche gloutonne

- Tirer une fois chaque bras. $\hat{\mu}_i = r \sim \nu_i$
- Au temps u, tirer le bras tel que

$$i_t = argmax\{\hat{\mu}_{i,t-1}, i = 1 \dots K\}$$

Un exemple

- 2 bras: bras 1, $\mu_1 = .8$; bras 2, $\mu_2 = .2$.
- Les deux premiers tirages donnent $r_1 = 0$ (pour le bras 1), et $r_2 = 1$ (pour le bras 2).
- Que se passe-t-il ?

The simplest approach: ϵ-greedy selection

At each time t,

- With probability $1 - \varepsilon$
  select the arm with best empirical reward

$$i_t = argmax\{\hat{\mu}_{i,t-1}, i = 1 \dots K\}$$

- Otherwise, select i_t uniformly in $\{1 \dots K\}$

Regret of ϵ-greedy selection

$$\mathbb{E}[Regret_T] > \varepsilon T \frac{1}{K} \sum_i \Delta_i$$

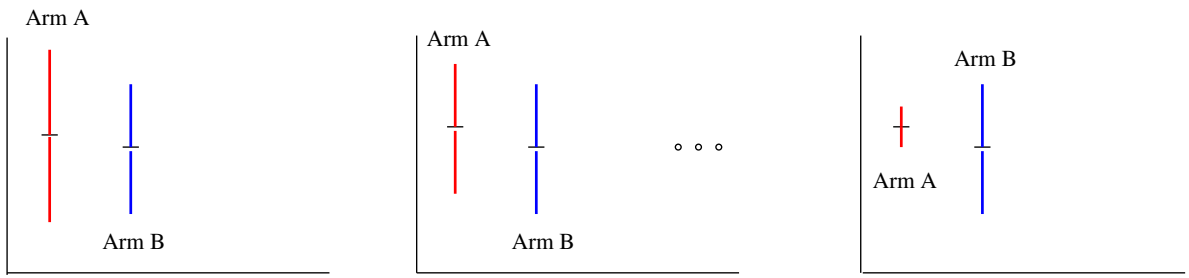
On a donc un regret linéaire en T...

Overview

Upper Confidence Bound

Auer et al. 2002

$$\text{Select } i_t = \text{argmax} \left\{ \hat{\mu}_{i,t} + \sqrt{\frac{3log(t)}{2n_{i,t}}}, i = 1 \dots K \right\}$$



Decision: Optimism in front of unknown !

Upper Confidence bound, 2

Thm: UCB achieves the optimal regret rate $log(t)$

$$\text{Select } i_t = \text{ argmax } \left\{ B(i,t) = \hat{\mu}_{i,t} + \sqrt{\frac{3log(t)}{2n_{i,t}}}, i = 1 \dots K \right\}$$

Hoeffding inequality

Soit un ensemble de tirages $r_1, \dots r_n$ d'une variable aléatoire à valeur dans $[0,1]$, d'espérance μ,
soit la moyenne empirique $\hat{\mu}_n = 1/n \sum_{u=1}^{n} r_u$, alors

$$\begin{aligned} \mathbb{P}\left(\hat{\mu}_n - \mu \geq \varepsilon\right) &\leq \exp\left(-2\,\varepsilon^2 n\right), \\ \mathbb{P}\left(\mu - \hat{\mu}_n \geq \varepsilon\right) &\leq \exp\left(-2\,\varepsilon^2 n\right), \\ \mathbb{P}\left(|\hat{\mu}_n - \mu| \geq \varepsilon\right) &\leq 2\exp\left(-2\,\varepsilon^2 n\right) \end{aligned}$$

Upper Confidence bound, 3

Par application de Hoeffding, on a:

$$\Pr\left(\hat{\mu}_{i,t} + \sqrt{\frac{3log(t)}{2n_{i,t}}} \leq \mu_i\right) \leq t^{-3}$$

et

$$\Pr\left(\hat{\mu}_{i,t} - \sqrt{\frac{3log(t)}{2n_{i,t}}} \geq \mu_i\right) \leq t^{-3}$$

Considérons les évènements au temps t, $\hat{\mu}_{i,t}$ est dans l'intervalle de confiance:

$$(a) : \hat{\mu}_{i,t} \leq \mu_i + \sqrt{\frac{3log(t)}{2n_{i,t}}}$$

et $\hat{\mu}_{*,t}$ aussi:

$$(b) : \mu_* - \sqrt{\frac{3log(t)}{2n_{*,t}}} \leq \hat{\mu}_{*,t}$$

Upper Confidence bound, 4

Si dans ces conditions le bras i est choisi au lieu du bras $*$, ca implique que:

$$\mu_i + 2\sqrt{\frac{3log(t)}{2n_{i,t}}} \geq \mu_*$$

donc

$$n_{i,t} \leq \frac{6\log(t)}{\Delta_i^2}$$

Bornons $n_{i,t}$

Soit u un entier. On a:

$$\begin{array}{rcl} n_{i,t} & \leq & u + \sum_{s=u+1}^{t} \mathbb{1}\{i_s = i \cap n_{i,s} > u\} \\ & \leq & u + \sum_{s=u+1}^{t} \mathbb{1}\{B(i,s) \geq B(*,s) \cap n_{i,s} > u\} \ . \end{array}$$

Or, $\{B(i,s) \geq B(*,s)\}$ implique que $n_{i,s} \leq \frac{6\log(s)}{\Delta_i^2}$

ou que (a) ou (b) est faux.

Upper Confidence bound, 5

Choisissons $u = \frac{6\log(t)}{\Delta_i^2}$. Alors ou bien (a) ou bien (b) est faux. Chacun de ces évènements arrive avec proba moins que t^{-3}. En prenant une borne union (la proba d'une union d'evenements est moins que la somme de la proba de ces évènements), on a

$$\begin{aligned} \mathbb{E}(n_{i,t}) &\leq \frac{6\log(t)}{\Delta_i^2} + \sum_{s=u+1}^{t} \left[\sum_{a=u+1}^{s} a^{-3} + \sum_{a=1}^{s} a^{-3} \right] \\ &\leq \frac{6\log(t)}{\Delta_i^2} + \frac{\pi^2}{3}. \end{aligned}$$