

Module Master Recherche

Apprentissage et Fouille

Michele Sebag — Balazs Kegl — Antoine Cornuéjols
<http://tao.lri.fr>

4 février 2008

Ce cours: Apprentissage de stratégies

1. Bandits manchots
2. Etude de cas: Application au jeu de Go

Bandits Manchots

Notations et hypothèses de travail

- ▶ K bras
- ▶ Bras i : probabilité p_i de gain $p_i \in [0, 1]$
- ▶ Gains de chaque bras: indépendants, identiquement distribués
- ▶ Bras indépendants

Stratégie

- ▶ A chaque temps t
- ▶ Un bras est choisi Critère C
- ▶ Disons le bras i ; on recoit un gain r_t tiré d'après p_i
- ▶ Score $:=$ Score $+ r_t$

Objectif

Maximiser le gain cumulé

Exploration vs Exploitation

Le dilemme

A tout moment :

- ▶ il y a les meilleurs bras connus $\rightarrow$ les jouer
- ▶ mais on a peut-être manqué LE MEILLEUR $\rightarrow$ jouer aussi les autres bras

Exploiter

Explorer

Minimiser le regret

- ▶ $p^* = \operatorname{argmax}\{p_i, i = 1 \dots K\}$
- ▶ Performance maximum après T pas de temps: $T p^*$
- ▶ Regret : Perf. maximale - Perf réelle

$$\mathcal{L}(T) = T p^* - \sum_{t=1}^T r_t$$

Le principe

Ayant joué un bras N fois avec les gains $r_1, \dots, r_N$

- ▶ On a une estimation de sa probabilité de gain

$$\hat{p}_N = \frac{1}{N} \sum_{t=1}^N r_t$$

- ▶ On connaît la confiance de cette estimation
Borne de Hoeffding: after N trials,

$$P(|\hat{p}_N - p| > \epsilon) < \exp\{-2N\epsilon^2\}$$

Le principe, 2

Optimisme en face de l'incertain

Au temps T , jouer le bras maximisant
moyenne empirique + confiance

$$\text{Select } i^* = \operatorname{argmax}\left\{\hat{p}_i + \sqrt{\frac{2\log T}{n_i}}, i = 1 \dots K\right\}$$

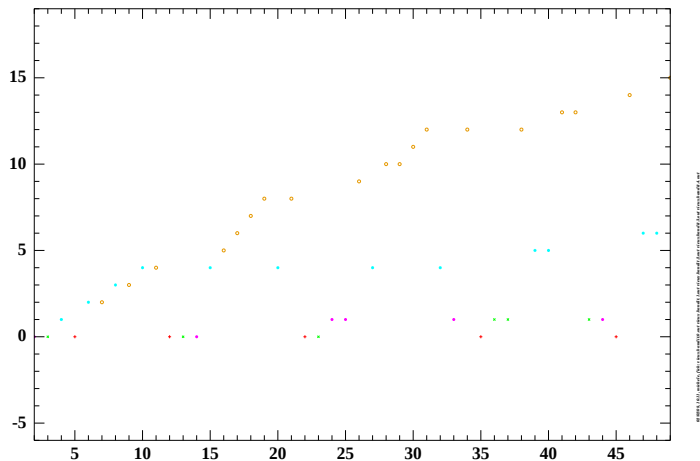
► Terme d'exploitation

$$\hat{p}_i$$

► Terme d'exploration

$$\sqrt{\frac{2\log T}{n_i}}$$

Upper Confidence Bound, 2



Upper Confidence Bound, 3



Variantes

Ajustement du terme d'exploration

$$\hat{p}_i + \sqrt{\frac{2 \log T}{n_{i,t}}} \quad \text{UCB1 Auer02}$$

$$\hat{p}_i + \frac{\log T}{n_{i,t}} + \sqrt{\frac{v_{i,t} \log T}{n_{i,t}}} \quad \text{UCB-Tuned Auer 02}$$

$$\hat{p}_i + \sqrt{\frac{v_{i,t} \log T}{n_{i,t}}} \quad \text{KUCBT Kocsis 06}$$

$$\hat{p}_i + \sqrt{\frac{c \log T}{n_{i,t}}} \quad \text{c-UCB}$$

Où $v_{i,t} = \max(\varepsilon, V_{i,t})$ et $V_{i,t}$ est la variance empirique du i -th bras.