

TEST A/B

INTRODUCTION À L'APPRENTISSAGE PAR RENFORCEMENT

Guillaume Wisniewski
guillaume.wisniewski@limsi.fr

Université Paris Sud — LIMSI

mars 2016

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

1 / 37

Première partie I

CADRE

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

2 / 37

COMMENT CHOISIR SA PIZZA ?



- J'aime beaucoup les « 4 fromages »
- Quelle pizza dois-je commander quand j'arrive dans un restaurant ?

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

3 / 37

COMMENT CHOISIR SA PIZZA ?



- J'aime beaucoup les « 4 fromages »
- Quelle pizza dois-je commander quand j'arrive dans un restaurant ?
- Est-ce que je dois tout le temps prendre une 4 fromage ?
au risque de rater quelque chose de très bon
- Est-ce que je dois goûter une autre pizza ?
au risque d'être toujours déçu

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

3 / 37

POUR LA CULTURE



Feynman's restaurant problem

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

4 / 37

DE MANIÈRE PLUS SÉRIEUSE...

Comment choisir sa page d'accueil ?

- la page d'accueil a potentiellement un impact fort sur les visiteurs
- trop de changements / changement de mauvais goût → non convaincant, perte d'utilisateurs, ...
- pas assez de changement → lassitude, page plus à la mode, ...

Évaluation : nombre d'achats / enregistrement d'utilisateur / ...

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

5 / 37

EXEMPLE N° 2

- Washington Post
- objectif : évaluer l'impact des modifications d'un articles voire de l'ajout d'un contenu
- formellement : plusieurs « variantes » du site; arriver à trouver la meilleure variante en fonction des retours des utilisateurs en continu

<https://developer.washingtonpost.com/pb/blog/post/2016/02/08/bandito-a-multi-armed-bandit-tool-for-content-testing/>

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

6 / 37

EXEMPLE DE RÉSULTATS (1)



William Merideth says he asserted his right to privacy and property by shooting down a drone. (Luke Sharrett for The Washington Post)

You may be powerless to stop a drone from hovering over your own yard

The fight over whether a Kentucky man had the right to shoot down a quadcopter flying over his property has weighty implications for the future of drones — and for the tech companies that dream of using them to deliver packages.

By Andrea Peterson and Matt McFarland



William Merideth asserts his right to privacy. (Luke Sharrett for The Washington Post)

You may be powerless to stop a drone from hovering over your own yard

The fight over whether a Kentucky man had the right to shoot down a quadcopter flying over his property has weighty implications for the future of drones.

By Andrea Peterson and Matt McFarland

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

6 / 37

EXEMPLE DE RÉSULTATS (1)



William Merideth says he asserted his right to privacy and property by shooting down a drone. (Luke Sharrett for The Washington Post)

You may be powerless to stop a drone from hovering over your own yard

The fight over whether a Kentucky man had the right to shoot down a quadcopter flying over his property has weighty implications for the future of drones — and for the tech companies that dream of using them to deliver packages.

By Andrea Peterson and Matt McFarland

CTR : 11,7%



William Merideth asserts his right to privacy. (Luke Sharrett for The Washington Post)

You may be powerless to stop a drone from hovering over your own yard

The fight over whether a Kentucky man had the right to shoot down a quadcopter flying over his property has weighty implications for the future of drones — and for the tech companies that dream of using them to deliver packages.

By Andrea Peterson and Matt McFarland

CTR : 24,6%

Évaluation : Click-Through Rate (taux de clics)

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

7 / 37

EXEMPLE DE RÉSULTATS (2)



Why Marie Kondo's life-changing magic doesn't work for parents

The celebrity organizer had her first baby in July, and I'd be willing to bet even she has slipped once or twice since then.

By Tanya C. Snyder • On Parenting

CTR : 3,3%



The real reasons Marie Kondo's life-changing magic doesn't work for parents

The celebrity organizer had her first baby in July, and I'd be willing to bet even she has slipped once or twice since then.

By Tanya C. Snyder • On Parenting

CTR : 3,9%



The real reasons Marie Kondo's life-changing magic doesn't work for parents

The celebrity organizer had her first baby in July, and I'd be willing to bet even she has slipped once or twice since then.

By Tanya C. Snyder • On Parenting

CTR : 4,8%

DIFFICULTÉ DE L'ÉVALUATION



- changement + observation d'une variation → aucune conclusion possible
- possibilité de « fluctuation locale » (ex. : le changement est fait juste avant/après Noël)

⇒ besoin d'une expérience contrôlée

DIFFICULTÉ DE L'ÉVALUATION



- changement + observation d'une variation → aucune conclusion possible
- possibilité de « fluctuation locale » (ex. : le changement est fait juste avant/après Noël)

⇒ besoin d'une expérience contrôlée = continuer d'observer ce qu'il se passe avec l'ancienne version

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

9 / 37

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

9 / 37

MAIS ENCORE...

- allocation de ressources en gestion de projet
- tests médicaux
- gestion de portefeuilles en bourse
- placement de publicités
- *adaptive routing*
- ...

Multi-Armed Bandit

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

10 / 37

IDÉE ESSENTIELLE



Ce n'est plus de l'apprentissage supervisé

- apprentissage supervisé : réponse optimale connue et on en déduit une récompense
- renforcement : on ne connaît que la récompense associée à la décision sans aucun moyen de connaître la récompense des autres choix

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

11 / 37

FORMALISATION



- *n* choix (= arms)
- le i^e choix donne une récompense de 1 avec une probabilité p_i fixe mais inconnue
- à l'instant t on choisit le choix a_t
- on obtient une récompense (= reward) $r_t \sim Ber(p_{a_t})$

Expérience répétée N fois \Rightarrow maximiser la récompense totale

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

12 / 37

LE DILEMME EXPLORATION/EXPLOITATION



- exploitation : utilise la connaissance accumulée pour prendre la bonne décision
- exploration : prend une décision « non optimale » pour mettre à jour ses connaissances

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

13 / 37

POURQUOI EST-CE UN DILEMME ?

- si l'on ne fait qu'explorer : risque de ne prendre que de mauvaises décisions \rightarrow récompense cumulée faible
- si l'on ne fait qu'exploiter : risque de rater de nouvelles opportunités + modification des récompenses au cours du temps (lassitude) \rightarrow performance cumulée non optimale

\Rightarrow trouver le bon compromis

\Rightarrow apprentissage *on-line* des préférences

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

14 / 37

UN PEU PLUS DE NOTATIONS

NOTATIONS

- H_t ensemble des informations que l'on a vues jusqu'à l'instant t (actions prises & récompenses reçues)
- un algorithme d'apprentissage π associe H_t à une distribution de probabilité sur l'ensemble des actions \mathcal{A}

CRITÈRE D'ÉVALUATION

$$\text{Regret}(T, \pi, f^*) = \mathbb{E} \left[\sum_{t=1}^T f^*(a^*) - f^*(a_t) \right] \quad (1)$$

- a_t action prise à l'instant t
- a^* action optimale à prendre à l'instant t
- f^* « vraie » récompense

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

15 / 37

Deuxième partie II

MÉTHODES

SIMULER UNE DISTRIBUTION DISCRÈTE EN PYTHON

Fonction qui retourne Exploration avec une probabilité ϵ et Exploitation avec une probabilité $1 - \epsilon$

```
import random
def discret_rand(epsilon):
    if random.random() > epsilon:
        return 0
    else:
        return 1
```

ÉTAPE D'EXPLORATION

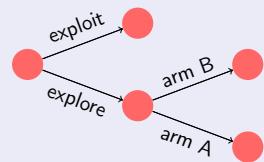


- prend la décision la plus « rentable »
 - « valeur » d'une décision :
- $$\frac{\sum \text{récompenses obtenues avec ce choix}}{\#\text{choix}}$$
- obtient une récompense
 - garde une « trace » de la décision qui a été prise et de la récompense associée comme avant

L'ALGORITHME ϵ -greedy

PRINCIPE

À chaque étape, choisi aléatoirement entre un choix « exploration » et un choix « exploitation » :



FORMELLEMENT

- hyper-paramètre de l'algorithme : ϵ
- avec une probabilité ϵ : exploration
- avec une probabilité $1 - \epsilon$: exploitation

ÉTAPE D'EXPLOITATION



- prend une décision au hasard
- obtient une récompense
- garde une « trace » de la décision qui a été prise et de la récompense associée => informations nécessaires pour prendre la phase d'exploitation

ÉTAPE D'EXPLOITATION

ALGORITHME ϵ -GREEDY

À chaque itération t :

- estimer $\hat{f}_t = \mathbb{E}[f^* | H_t]$
- avec une probabilité ϵ choisir $a_t \sim \text{Unif}(\mathcal{A})$, sinon choisir $a_t = \arg \max_a \hat{f}_t$

GARANTIES THÉORIQUES

- si ϵ est suffisamment petit et décroît avec t , l'algorithme ϵ -greedy converge vers la solution optimale $\text{Regret}(T, \pi)$ en un temps exponentiel.

VOCABULAIRE : LIEN AVEC TEST A/B

- ϵ -greedy : algorithme développé dans le cadre de l'apprentissage par renforcement (définition à suivre)
- test A/B : même principe, dans la communauté « marketing »
 - ▶ en général, uniquement de l'exploitation après N décisions
 - ▶ moins efficace (à tester)

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

22 / 37

Troisième partie III

ÉVALUATION D'UN ALGORITHME DE BANDIT

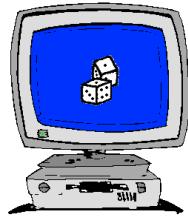
G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

23 / 37

COMMENT TESTER



- aspect le plus compliqué des algorithmes de bandit
- application : suite de décisions \Rightarrow plus de corpus d'apprentissage / test
- manière de tester = simulation = méthode de Monte-Carlo [Metropolis, 47].
[Ulam et von Neumann]

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

24 / 37

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

25 / 37

EXEMPLE

2 PIZZAS

Choix entre une pizza « 4 fromages » et une « reine » :

- chaque pizza est représentée par un bras
- simulation \rightarrow on connaît la bonne réponse (mais on l' « oublie ») :
 - ▶ bras « 4 fromage » : récompense avec une probabilité de 0,6
 - ▶ bras « reine » : récompense avec une probabilité de 0,2
- raffinement : récompense différente pour chaque pizza

CAS GÉNÉRAL

- autant de bras que de pizza
- techniquement : liste de BernoulliArm

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

26 / 37

PRINCIPE DE LA SIMULATION (1) : LES CHOIX

PRINCIPE

- k choix
- chaque choix permet d'obtenir une récompense de 1 avec une probabilité p et de 0 avec une probabilité $1 - p$
- techniquement : Bernoulli Arm

EN PYTHON (1 BRAS)

```
class BernoulliArm:  
    def __init__(self, p):  
        self.p = p  
  
    def draw(self):  
        return 0.0 if random.random() > self.p else 1.0
```

G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

25 / 37

PRINCIPE DE LA SIMULATION (2) : LA BOUCLE

- on « joue » pendant n tours (= horizon)
 - ① on prend un décision
 - ② on obtient une récompense
 - ③ on met à jour ses connaissances
- qualité de l'algorithme : récompense à chaque tour *ou* récompense cumulée

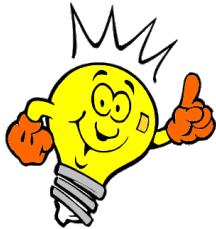
G. WISNIEWSKI (UPS)

TEST A/B

MARS 2016

27 / 37

ASTUCE



- résultat de la simulation = dépend du hasard
- les résultats peuvent varier grandement d'une expérience à l'autre
 - ▶ surtout si l'on répète l'expérience jusqu'à avoir un gain
- solution : répéter la simulation k fois \oplus moyennez la récompense à chaque instant sur l'ensemble des simulations

AU FINAL

```
def test_algorithm(algo, means, num_sims, horizon):
    # init. all decisions
    arms = [BernoulliArm(mu) for mu in means]
    rewards = []

    for sim in range(num_sims):
        algo.initialize(len(arms))

    for t in range(horizon):
        chosen_arm = algo.select_arm()
        reward = arms[chosen_arm].draw()
        algo.update(chosen_arm, reward)

    rewards.append(reward)

return np.array(rewards).reshape((num_sims, horizon))
```

INTERFACE D'UN ALGORITHME A/B

```
class EpsilonGreedy:

    # constructor
    # epsilon (float): tradeoff exploration/exploitation
    def __init__(self, epsilon): pass

    # re-initialize the algorithm in order to run a new simulation
    # n_arms (int): number of arms
    def initialize(self, n_arms): pass

    # return a index of the chosen decision
    def select_arm(): pass

    # update knowledge
    # chosen_arm (int): the decision that has been made
    # reward (float): the obtained reward
    def update(self, chosen_arm, reward): pass
```

Quatrième partie IV

APPRENTISSAGE PAR RENFORCEMENT

APPRENTISSAGE PAR RENFORCEMENT

- apprentissage d'une suite d'**actions** de manière à optimiser une récompense numérique au cours du temps
- formalisation :
 - ▶ un ensemble d'états S modélisant l'environnement dans lequel un agent évolue p.ex. : pièces d'un labyrinthe
 - ▶ un ensemble d'actions A que peut effectuer l'agent p.ex : différent type de déplacement
 - ▶ un superviseur capable de donner une récompense à chaque actions / suite d'action p.ex. : récompense quand on a trouvé une sortie

APPLICATIONS



C'est « juste » :

- du ϵ -greedy
- avec une meilleure estimation des récompenses liées aux actions (apprentissage supervisé)
- une très grosse puissance de calcul

Cinquième partie V

TP

À VOTRE TOUR !



1^{RE} PARTIE

- ❶ On note r_t la qualité d'une action tel qu'estimée à l'instant t .
Comment calculer r_t
- ❷ Exprimer r_{t+1} en fonction de r_t .
- ❸ Quel est l'intérêt de la relation précédente ?
- ❹ Observer l'évolution de la récompense à chaque instant pour les paramètres :
 - ▶ $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$
 - ▶ horizon : 250
 - ▶ nombre de simulations : 100
 - ▶ arms :

```
means = [0.1, 0.1, 0.1, 0.1, 0.9]
random.shuffle(means)
arms = [BernoulliArm(mu) for mu in means]
```
- ❺ concluez !

2^E PARTIE : RAFFINEMENT

- ❶ Limite de l'algorithme ϵ -greedy : choix des actions non informées. Que se passe-t-il quand A a une récompense dans 99% des cas et B dans seulement 10% des cas ?
- ❷ Algorithme SoftMax : si on a n choix, on choisi le i^e avec une probabilité :
$$\frac{e^{\frac{r_i}{\tau}}}{\sum_{j=1}^n e^{\frac{r_j}{\tau}}}$$
où τ température (hyper-paramètre). Implémentez cette stratégie et observez son comportement pour différentes valeurs de τ .
- ❸ Raffinement : annealing = limiter les étapes d'exploration au fur et à mesure que l'on accumule de l'information :
$$\tau = \frac{1}{t + 10^{-7}}$$
Quel est l'intérêt de ce choix (expérimentalement et « intuitivement ») ?