Projet Exploration/Exploitation

Ryan Pierre Fumeron

Table des matières

Introduction

Dans ce projet, nous nous sommes intéressés au dilemme de l'exploration vs exploitation. Prenons l'exemple du bandit manchot pour illustrer ce problème : Imaginons plusieurs machines à sous, chaque machine a un gain et une probabilité de toucher ce gain qui lui sont propres. Un joueur, qui veut maximiser son profit, a donc plusieurs options : jouer constamment la même machine sans acquérir d'informations sur les autres et donc sans savoir s'il y a une meilleure solution. Ou bien, jouer toutes les machines pour en savoir plus sur leurs probabilités mais en gagnant moins d'argent que si on avait uniquement joué la "meilleure" machine (celle qui offre le meilleur profit). Ce problème trouve de nombreuses applications en intelligence artificielle mais également dans la vie quotidienne, par exemple : Vous êtes dans un restaurant, commandez-vous un plat que vous appréciez ou bien un plat que vous n'avez jamais goûté en espérant qu'il soit meilleur que celui que vous connaissez ?
Nous avons donc étudié des algorithmes du dilemme exploration vs exploitation avec 2 jeux différents. Premièrement, nous nous sommes intéressés à l'exemple du bandit-manchot décrit ci-dessus, avec comme objectif de maximiser le gain. Il faut pour cela trouver un bon compromis entre l'exploration des différentes machines puis l'exploitation de la machine la plus rentable.
Le deuxième jeu est celui du morpion où nous avaons fait s'affronter différents algorithmes : aléatoire, UCT et Monte-Carlo).

1 Bandits-Manchots

Supposons une machine à sous à N leviers dénotés par l'ensemble 1, 2, . . . , N. Chacun de ses leviers est une action possible parmi lesquelles le joueur doit choisir à chaque pas de temps : l'action choisie à l'instant t sera appelée at (un entier entre 1 et N). Nous supposons que la récompense associée à chaque levier i suit une distribution de Bernoulli de paramètre μi : $avecuneprobabilitilejoueurobtientunercompensede1, avecunep$
$-direquelesisontconstantstoutaulongdelapartie.Pourlejoueur, legainauboutdeTpartiesestlasomm$
$PTt = 0rt(n.b.larcompensetantalatoire, legainGTestunevariablealatoire, toutcommert).Sonbutes$
$i = argmaxi1, ..., Nietlerendementassoci := i = maxi1, ..., ni.Silejoueurjoueunautrelevierquei, il$
$PTi = 1rt, avecrtlarcompensealatoiretiredeladistributiondeBernoullideparamtre.Onappelleregre$
$LT = GPTt = 1rt = PTt = 1(rtrt).L'objectifestdoncdeminimiserceregret.$

Question 1