

Apprentissage par renforcement : contrôle de stock

Aurélien Garivier

27 octobre 2011

Le but de ce TP est de contrôler le stock d'un vendeur de machines à laver de manière à optimiser son profit. Le cadre est le suivant :

- on note X_t le nombre de machines à laver qu'il détient dans son magasin à la fin de la semaine numéro t ;
- au maximum, il peut stocker M machines ; mais chaque semaine, le coût d'entretien et de lavage de chaque machine est de h ;
- il peut acheter A_t machines au prix unitaire c : il les reçoit alors au début de la semaine suivante ; les frais de livraison sont de K quel que soit le nombre de machines livrées, sauf bien sûr si $A_t = 0$;
- il vend ses machines au prix p ; le nombre de clients au cours de la semaine t est une variable aléatoire D_t , et on suppose que la suite $(D_t)_{t \geq 1}$ est i.i.d.
- on note R_t son chiffre d'affaire sur la semaine t ;
- on suppose qu'au temps initial $t = 1$, son stock est plein : $X_1 = M$;
- soit $\gamma \in]0, 1[$ tel que le taux d'inflation par semaine soit égal à $\gamma^{-1} - 1$: on suppose donc que le vendeur cherche à maximiser en espérance son profit actualisé

$$\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \right].$$

On pourra prendre, par exemple : $M = 15$, $K = 0.8$, $c = 0.5$, $h = 0.3$, $p = 1$, $\gamma = 0.99$ et pour la distribution de D_t une loi géométrique de paramètre 0.1.

Simulation

Montrer que

$$X_{t+1} = \left((X_t + A_t) \wedge M - D_{t+1} \right)_+,$$

$$R_{t+1} = -K \mathbb{1}\{A_t > 0\} - c \left((X_t + A_t) \wedge M - X_t \right)_+ - h X_t + p \left(D_{t+1} \wedge (X_t + A_t) \wedge M \right).$$

Comment évolue son chiffre d'affaire s'il commande chaque semaine exactement deux machines ?

Paramètres du MDP

1. Pour chaque $(x, y, a) \in \{0, \dots, M\}^3$, calculer $P(X_{t+1} = y | X_t = x, A_t = a)$. Construire en R une liste `trans` telle que

$$\text{trans}[[\mathbf{a}]][\mathbf{x}, \mathbf{y}] = P(X_{t+1} = y | X_t = x, A_t = a)$$

2. Pour chaque $(x, a) \in \{0, \dots, M\}^2$, calculer $\mathbb{E}[R_{t+1} | X_t = x, A_t = a]$. Construire en R une liste `rew` telle que

$$\text{rew}[[\mathbf{a}]][\mathbf{x}] = \mathbb{E}[R_{t+1} | X_t = x, A_t = a]$$

Evaluation d'une politique

On code une politique `pol` par un tableau tel que `pol[1+x]` désigne le nombre de machines à acheter quand $X_t = x$. Ecrire une fonction `policyValue <- function(pol)` qui calcule la fonction valeur d'une politique `pol`.

Que peut-il espérer gagner en achetant exactement deux machines chaque semaine ?

Itération sur les valeurs

Programmer l'opérateur de Bellman de sorte que la fonction `BellmanOperator <- function(V)` renvoie $\text{res}\$V = T^*(V)$ ainsi que la politique gloutonne associée `res$pol`.

Programmer l'algorithme d'itération sur les valeurs pour trouver la stratégie optimale pour le vendeur.

Itération sur les politiques

Programmer l'algorithme d'itération sur les politiques, et commenter le résultat. En particulier, on comparera le temps de calcul avec l'algorithme d'itération sur les valeurs.

Q-learning

On suppose maintenant que le vendeur ne connaît pas la loi de D_t . Implémenter la méthode de Q-learning, et vérifier sa convergence quand t tend vers l'infini.

Question théorique

Quelle est la forme que l'on trouve pour la politique optimale ? Changer les paramètres du problème, pour voir si l'on trouve toujours une politique semblable. Prouver ce résultat, sous des hypothèses appropriées.