*Réunion 21/06/2011*

**L’algorithme de l’apprentissage par renforcement**

1. **Les définitions de base**

Dans mon stage, l’algorithme de l’apprentissage par renforcement utilisé est ***SARSA*** qui correspond à des mots anglais ***State-Action-Reward-State-Action*** et des mots français ***Etat-Action-Récompense–Etat-Action***.

Ainsi, il faut clairement montrer l’ensemble d’Etats, l’ensemble d’Actions et la formule qui calcule la valeur de récompense quand on passe de l’état St à l’état St+1



Pour montrer ces ensembles, d’abord on doit répondre des questions suivantes :

1. **Qu’est ce qu’un état**

Un état noté est S qui décrit les nombres de personnes dans chaque groupes s, e, i, r par ville comme suivant :

*S = ((s1,e1,i1,r1),(s2,e2,i2,r2),….,(sn,en,in,rn))*

Avec : *n* = nombre de villes

(si,ei,ii,ri) ϵ R4

1. **Abstraction d’état**

Au moment *t*, on a l’état St et ensuite on doit passer à l’état St+1

1. **Est-ce que l’on garde s, i seulement ou i seulement**

Comme au-dessus, on définit un état qui contient quatre groupes de personnes *s, e, i, r*. Cependant, cela rend un état plus complexe. Alors, on va définir un état qui contient *s, i* seulement ou *i* seulement. Parce que, ce qui nous intéresse est les nombres de personnes sensibles et de personnes infectées.

De plus, il est très important pour la décision « garder *s, i* seulement ou *i* seulement » parce que cela influence beaucoup le nombre d’états.

Si on garde *i* seulement, on a un état Sa1= (i1, i2, i3,…, in)

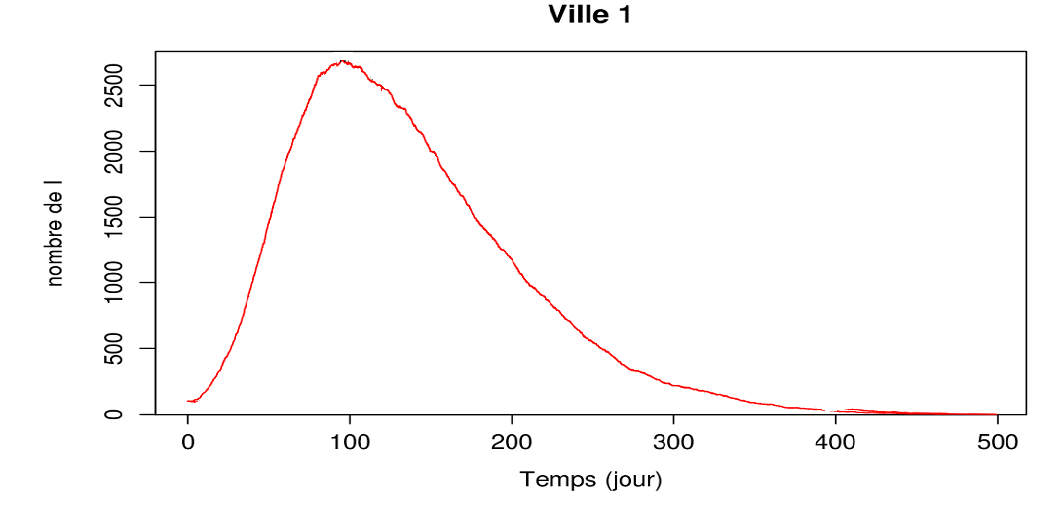
Si on garde *s, i* seulement, on a un état Sa2= ((s1, i1), (s2, i2), (s3, i3),…, (sn,in)).

*Alors, pour simplifier plus, d’abord on va garder i seulement.*

1. **Discrétiser les valeurs de s, e, i, r**

Car on garde seulement les valeurs du nombre de personnes infectées, alors on va discrétiser la valeur de i.

* D’abord, on choisit le nombre de villes.
* Après, on fonctionne le programme de stage sans vaccination avec les valeurs de paramètres.
* Ensuite, on choisit une chiffre ***kDiscret*** que on va utiliser pour discrétiser la valeur de I.
* Enfin, on discrétise de façon uniforme les valeurs de I par ville, par exemple pour la première ville sans vaccination, *kDiscret* = 10, alors on récupère 10 valeurs discrètes de I. Cependant, comme on voit dans la courbe de la ville, il y a deux parties, une partie monte et l’autre descend. Ainsi, on donne une signe qui a deux symboles + ou – (+ est « *monter* », - est « *descendre* »)
* Alors, on récupère 10\*2 = 20 valeurs de I qui correspondent aux valeurs de s, e, r.



Enfin, un état S au moment *t* est St qui est défini comme suivant :

*St=(I1,signeI1, I2, signeI2,…, In, signeIn)*

Avec Ii a *kDiscret* valeurs correspondantes

signeIi est (+ ou -)

1. **Nombre d’états**

D’abord, on a le nombre de villes, par exemple *nbVilles*= 3

Ii a *kDiscret* valeurs correspondantes, *kDiscret* =10

signeIi a deux symboles +/-

Alors, le nombre d’état = (kDiscret \*2)nbVilles

= (10\*2) 3

= 8000 (états)

*Pour chaque état, on a une action pour chaque ville concernant à vacciner*.

1. **Discrétiser les actions**

* D’abord, pour l'instant on vaccine toutes les villes. Concernant le "quand" pour l'instant on choisit de vacciner à des instants précis *tstart + k \* T*

*(k= 0->KMAX=INT((tmax -tstart)/T))* étant donné tstart et tmax.

* Ensuite, considérons que l’on a au départ *V* vaccins. On définit l'unité de vaccination comme étant *UV= V / KMAX*.
* Au départ *KPOS* (K possible) vaut *KMAX*.
* On va déjà pour l'instant apprendre quelles sont les stratégies optimales pour vacciner à chaque pas entre 0 et KPOS.
* Une stratégie est par exemple de faire (1, 1, 1,1,....,1)

c.à.d. de vacciner 1 UV à chacun des KMAX temps. Une autre est de faire (KMAX, 0, 0,0,...0). Par exemple, si KMAX=3 on a

|  |  |
| --- | --- |
| * + 0 0 3   + 0 1 2   + 0 2 1   + 0 3 0   + 1 0 2 | * + 1 1 1   + 1 2 0   + 2 0 1   + 2 1 0   + 3 0 0 |

* + Ce nombre est le nombre d'élément dont la somme des termes vaut (2KMAX-1 KMAX)

soit (2KMAX-1)\*(2KMAX-2)....(KMAX)/KMAX!=(2KMAX-1)!/KMAX\*(KMAX-1)!^2

Si KMAX=3, on a 5\*4\*3/3\*2 =10 (élément)

* Enfin, on définit une valeur v.

*v ϵN, v ϵ{0, 1, 2, …, KMAX} \* UV*

Dans un état s, les actions possibles sont :

+ pour ville 1, on vaccine v1 personnes sensibles, S R ( S1, le nombre de personnes sensibles de la ville 1)

+ pour ville 2, on vaccine v2 personnes sensibles, S R ( S2, le nombre de personnes sensibles de la ville 2)

….

*Alors, une action a qui est décrit a(v1,v2,…,vn)*

1. **Combien d’action**

On a le nombre de villes, par exemple *nbVilles*= 3

On a KMAX fois de vacciner, *KMAX* =5

Alors, le nombre d’action  = KMAX nbVilles

= 53

= 125 (action)

*Pour chaque état, on a une action pour chaque ville concernant à vacciner*

1. **L’algorithme de l’apprentissage par renforcement**
2. **Introduction : les notations de base**

* Temps discret : t = tstart + k\*T

Avec k : 0KMAX

T : période de vaccination

* Etats : st ϵ S
* Actions : at ϵA(st)
* Récompenses : rt ϵ R(st)
* L’agent : st at
* L’environnement : (st,at) st+1, rt+1
* Politique : Πt : S A avec l’ensemble T, R
* Fonction d’évaluation Q(s,a) (*Q* c.à.d. qualité)

1. **L’algorithme de l’apprentissage par renforcement**

s : état actuel

a : action actuelle

s’ : état suivant

a’ : action suivante

r(t+T)= - I - ∑vi ( i : 1 nbVilles)

I = I(t+T) – I(t)

 : taux d’apprentissage

  =

 : taux discount (en anglais, discounted rate),   = 1

 : taux d’exploitation

Avec la probabilité (1-l’action choisie *a* = argmax(Q(s,a))

Inversement, l’action choisie *a* = random (a)

En général, ou 5%

|  |
| --- |
| **début**  Initialiser *Q(s, a)* arbitrairement.  **répéter**  **pour chaque** *état* initial choisi pour **faire**  Initialiser *s*  Choisir l’action *a* par une procédure -gloutonne dérivée des valeurs de *Q(s, a)*  **répéter**  **pour chaque** *étape de l’état* **faire**  Exécuter l’action *a* ; recevoir le renforcement *r* ; et mesurer l’état suivant *s*’  Choisir l’action *a’* à partir de *s’* en utilisant une procédure -gloutonne dérivée des valeurs de *Q (s, a)*  ***Qnew (s, a)* Qold(s, a)+ [r + Qold(s’,a’) - Qold(s, a)]**  **s s’ ; a a’**  **fin**  **jusqu’à** *s est terminal*  **fin**  **jusqu’à** *critère d’arrêt* (convergence suffisante)  **fin** |

1. **Résultat après l’apprentissage par renforcement**

Avec le nombre de l’ensemble d’état est *nbEtat*, par exemple *nbEtat* = 8000.

Le nombre de l’ensemble d’action est *nbAction*, par exemple *nbAction* = 125

Alors, on a un dictionnaire pour les états avec chaque action ou une table

8000\*125 = 1000000.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | ……………….. |  |  |  |  |
| .  .  125 actions  . |  |  | ………………… |  |  |  |  |
|  |  | …………………. |  |  |  |  |
|  |  |  | …………….. |  |  |  |  |

8000 états