Reinforcement Learning

Master Recherche AIC LRI

14:00-16:00 Feb. 6th, 2017

Michèle Sebag & Diviyan Kalainathan

Documents are allowed. Computers are not allowed.

Read the exam document before you start. Any answer must be justified.
Il est conseillé de lire tout le texte avant de commencer. Les réponses doivent être justifiées.

1 Partie I. Questions de cours (7 points)

Q. 1.1 *Give the usual formalization of a reinforcement learning problem.*
Donnez la formalisation usuelle d'un probleme de renforcement.

Q. 1.2 *Give an example of non-Markovian decision process problem. Define what would be a policy for this problem ? Which algorithm would you use to build such a policy ?*
Donnez un exemple de problème de décision non markovien. Que serait la politique attendue dans ce cas ? Voyez-vous des algorithmes appropriés pour construire cette politique ?

Q. 1.3 *Describe* Policy iteration *and* Value iteration. *What are their main differences ? Discuss their strengths and weaknesses.*
Deux algorithmes de programmation dynamique sont *Policy iteration* et *Value iteration*. Quelles sont leurs principales différences ? Discutez leurs forces et faiblesses.

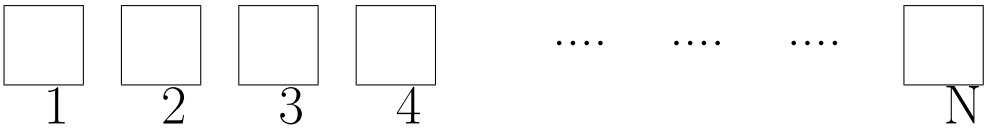
Q. 1.4 *In continuous state spaces, what does the "smoothness assumption" mean ? How does it help resolving the issue of dealing with an infinite amount of states?*
Dans les espaces d'état continus, que veut dire l'hypothèse de continuité (smoothness) ? Comment cette hypothèse aide-t-elle à résoudre l'apprentissage par renforcement ?

Q. 1.5 *How can neural nets be used for reinforcement learning ? What is learned ? What are the requisites ? Cite and discuss a few heuristics used to train them.*
Comment les réseaux neuronaux peuvent-ils être utilisés pour l'apprentissage par renforcement ? Quels sont les pré-requis ? Citez et discutez quelques heuristiques utilisées our les entrainer.

2 Partie II. Problem (8 points)

In the RiverSwim problem, the state space is the 1 × N gridworld. The set of actions is $\{Right, Left\}$. The N-th state is a terminal state. The reward function is $r(s) = 100$ iff $s = N$; otherwise $r(s) = 0$. The discount factor is $\gamma = .9$.
Dans le RiverSwim probleme, l'espace d'état est $\{1, \dots N\}$. *L'espace d'actions est* $\{Right, Left\}$. *L'état*

N est terminal. La récompense est de 100 dans l'état N, et de 0 partout ailleurs. Le facteur d'escompte est $\gamma = .9$.



Q. 2.1 *Let π be a random policy ($\pi(s) = Left$ or $Right$ with probability 1/2). Let the transition function be defined as:*

$$p(i, Right, i+1) = 1 \quad if \;\; i < N$$
$$p(i, Left, i-1) = 1 \quad if\; i > 0 \qquad else\; p(0, Left, 0) = 1$$

What is the probability of arriving in state N after 100 time steps ? After 1,000 time steps ?
Soit π une politique aléatoire ($\pi(s) = Left$ ou $Right$ avec probabilité 1/2). Avec la fonction de transition ci-dessus, quelle est la probabilité d'arriver à l'état N en 100 pas de temps ? en 1,000 pas de temps ?

Q. 2.2 *Let π be the constant policy, $\pi(s) = Right$ for all s. With same function transition as above, compute the value function (function of N).*
Soit π la politique constante qui va toujours à droite. Avec la meme fonction de transition que pour la question précédente, calculez la fonction de valeur associée à π (fonction de N).

Q. 2.3 *Same question with $p(i, Right, i+1) = .9;\; p(i, Right, i) = .1$.*
Même question avec une fonction de transition probabiliste, $p(i, Right, i+1) = .9;\; p(i, Right, i) = .1$.

3 Partie III. Apprentissage par démonstration (7 points)

Soit un robot équipé d'une camera à 512 pixels, de 10 capteurs infra-rouge à valeurs dans [0..255] et de 2 moteurs (vitesse de la roue gauche et droite). The problem goal is to build a patrolling controller.
Let a robot be equipped with a 512 pixel camera, 10 infra-red sensors ranging in [0..255] and 2 actuators (right and left wheel speed control). Le but est de construire une politique qui patrouille dans l'arène.

Q. 3.1 *What is the state space and the action space ?*
Quel est l'espace d'états et l'espace d'actions ?

Let a trajectory be given as $\{(s_t, a_t), t = 1 \dots T\}$ with s_t the robot state at time t and a_t the robot action selected by the human teacher at time t.
Soit une trajectoire $\{(s_t, a_t), t = 1 \dots T\}$ où s_t est l'état du robot à l'instant t et a_t l'action choisie par l'instructeur opérant le robot à l'instant t.

Q. 3.2 *Can you use a supervised learning algorithm to learn a policy from an ensemble of such trajectories ? Which algorithm would you use ? Why ? What might be the drawbacks of this controller ? What are the assumptions on the trajectories required to prevent these drawbacks ?*
Comment utiliser un algorithme d'apprentissage supervisé pour apprendre une politique à partir d'un ensemble de trajectoires ? Quel algorithme d'apprentissage utiliseriez-vous et pourquoi ? Quels sont les défauts possibles ou probables de cette politique ? Comment faudrait-il que les trajectoires soient générées pour apprendre une politique de bonne qualité ?x

Q. 3.3 *Assume the robot camera is a high-definition one (4096 pixels). Does that make the policy learning easier ? more difficult ? no impact ? What is the impact on the quality of the learned policy ? Why ?*
La camera du robot est remplacée par une camera haute définition (4096 pixels). Le problème d'apprendre une politique est-il plus facile ? plus difficile ? pareil ? Quel est l'impact sur la qualité de la politique apprise ? Pourquoi ?