Contrôle de Connaissance
Master Recherche Information, Apprentissage, Cognition - Université Paris-Sud 11
OPT 6, Robotique et Agents Autonomes

Vendredi 21 février 2014
Durée : 3h00
Documents autorisés : supports et notes de cours

Les exercices sont indépendants. Il est conseillé de lire tous les exercices avant de commencer. Rédigez les différentes parties sur des copies différentes pour faciliter la correction. Merci.

Partie I

1 Questions de cours (4 points)

Q. 1.1 *Qu'est-ce qu'un contrôleur de robot ? si c'était un programme, quel est son input, quel est son output ?*

Q. 1.2 *Quel était le problème d'apprentissage à résoudre dans le contexte du challenge DARPA 2005 ? (rappel, parcourir une route dans un désert, en allant le plus vite possible d'un point à un autre).*

Q. 1.3 *Supposons que l'on puisse guider le robot, en lui donnant des triplets (s, a, s') : dans l'état s, il est bon de faire l'action a, et on se retrouve alors (toujours ou souvent) dans l'état s'.*
Comment utiliser l'apprentissage supervisé classique pour apprendre un contrôleur ?
Quelles sont les limitations ?

Q. 1.4 *Quels sont les avantages et les inconvénients d'apprendre un contrôleur de robot en simulation, par rapport au fait de l'apprendre sur le robot physique ?*

2 Apprentissage par renforcement (8 points)

Considérons un processus de décision de Markov, décrit par : un espace d'états $\mathcal{S}$, un espace d'actions $\mathcal{A}$, un modèle de transition $p(s, a, s')$ défini sur $\mathcal{S} \times \mathcal{A} \times \mathcal{S}$, où $p(s, a, s')$ est la probabilité d'arriver dans l'état s' en choisissant l'action a dans l'état s, une fonction de récompense (reward) r définie sur $\mathcal{S}$, à valeurs dans $\mathbb{R}$ et bornée.

Etant donné une politique $\pi : \mathcal{S} \mapsto \mathcal{A}$, la fonction de valeurs V^π définie sur $\mathcal{S}$ et à valeurs dans $\mathbb{R}$, est l'espérance de la récompense cumulée en suivant la politique π à partir de l'état s, donnée par :

$$V_\pi(s) = \mathbb{E}_\pi\left[r(s) + \gamma r(s_1) + \gamma^2 r(s_2) + \ldots\right]$$

où l'espérance est prise sur les séquences d'etats $s_1, s_2, ..$ générées en tirant s_1 selon $p(s, \pi(s), \cdot)$, s_2 selon $p(s_1, \pi(s_1), \cdot)$, etc. Le facteur $0 < \gamma < 1$ est un facteur de discount (une récompense r vaut moins cher si elle est reçue plus tard).

Q. 2.1 *Quel est l'intérêt du facteur γ ?*
Indication : que se passe-t-il algorithmiquement si $\gamma = 1$?

La performance de la politique π est $V(s_0)$ si s_0 est l'état initial.
On admettra les équations de Bellman :

$$V_\pi(s) = r(s) + \gamma \sum_{s'} p(s, \pi(s), s')V(s')$$

$$Q_\pi(s, a) = r(s) + \gamma \sum_{s'} p(s, a, s')V(s')$$

On définit

$$Q^*(s, a) = max_\pi Q_\pi(s, a)$$

Q. 2.2 *Montrer que π est une politique optimale si et seulement si*

$$\pi(s) \in argmax\{Q^*(s,a), a\ in\ \mathcal{A}\}$$

Considérons tout d'abord le cas d'un espace d'états fini de taille S ($S = |\mathcal{S}|$). Considérer la matrice P_a de dimension $S \times S$ définie par $P_a(s_i, s_j) = p(s_i, a, s_j)$.

Q. 2.3 *Montrer que ses valeurs propres sont de module égal à 1. En déduire que $I - \gamma P_a$ est inversible, où I est la matrice identité de dimension $S \times S$, et $\gamma < 1$.*

Indication : Prenez un vecteur simple, dont une seule composante est égale à 1, et les autres à 0. Que voudrait dire le fait que ce vecteur soit un vecteur propre associé à une valeur propre > 1 ?

Q. 2.4 *Ecrire les équations de Bellman sous forme matricielle, en considérant Q_π comme un vecteur de dimension $S \times A$, avec A la taille de l'espace d'actions ($A = |\mathcal{A}|$).*

Q. 2.5 *Soit π une politique, associant à l'état initial s une action a. Montrer que π est optimal si et seulement si pour toute action $a' \neq a$*

$$(P_a - P_{a'})(I - \gamma P_a)^{-1} R \geq 0 \qquad (1)$$

où R est le vecteur de dimension S donné par $R = (r(s_1), \ldots r(s_S))$

2.1 Apprentissage par renforcement inverse

On suppose qu'on observe une politique optimale π^* (le comportement de l'expert). On se pose la question d'inférer la fonction de récompense r^* qui induit π^* comme politique optimale. L'équation (1) permet-elle d'inférer r^* ? (On pourra discuter l'unicité de r^*).

Notons a^*_s l'action $\pi^*(s)$. On veut maximiser

$$Q(s, a^*_s) - max_{a \neq a^*_s} Q(s,a)$$

On cherche à favoriser les fonctions de récompenses "simples".

Q. 2.6 *Quel terme de pénalisation sur r ajoute-t-on au problème de maximisation ci-dessus ? Discuter.*
Exprimer le problème d'apprentissage par renforcement inverse comme un problème d'optimisation sous contrainte.
Indication : s'inspirer de la formulation des Support Vector Machines.

2.2 Cas d'espaces infinis

L'approche ci-dessus ne marche plus si on considère un espace d'actions ou un espace d'états infini. Dans ce cas, on cherche la fonction de récompense r comme la combinaison linéaire d'un nombre fini de fonctions ϕ_i définies sur $\mathcal{S}$:

$$r(s) = \sum_{i=1}^{K} a_i \phi_i(s)$$

Q. 2.7 *Comment se ramener au cas précédent ? (Indication : définir les fonctions de valeurs $V_{i,\pi}$ associées à la fonction de récompense $r(s) = \phi_i(s)$ et utiliser le fait que*

$$\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$$