vinci
HAUTE ÉCOLE LÉONARD DE VINCI

Apprentissage par renforcement

Règle Vrai ou faux : Une bonne réponse vaut +1, une mauvaise réponse -1, aucune réponse 0. Pas de pénalité pour les autres questions.

Nom : _____

1. Vrai ou faux :
   (a) L'algorithme d'itération de la valeur trouve toujours vers la politique optimale quand utilisée jusqu'à convergence;
   VRAi : cf cours
   (b) Le principe d'optimalité de Bellman dit que je peux trouver une trajectoire optimale pour un problème alors qu'un des ses sous-problèmes ne peut être résolu de manière optimale;
   FAUX : traiter les ss-pb de manière optimale rend le problème original optimal
   (c) La fonction Q permet d'apprendre par essai-erreur;
   VRAi : cf cours
   (d) La Q-learning ne demande aucune connaissance préalable de l'environnement;
   VRAi : cf cours
   (e) Une instance MDP avec un discount factor γ petit tend à privilégier les récompenses à court terme;
   V : si γ vaut 0, on considère que les récompenses immédiates
   (f) $V(s)$ est de plus grande dimension que $Q(s,a)$ pour un même problème;
   FAUX : $|V| = |S|$ $|Q| = |S| \times |A|$
   (g) Pour maximiser sa récompense, il faut progressivement diminuer le facteur d'exploration ϵ au fur et à mesure que la politique s'affine;
   VRAi : si on explore trop on ne profite pas de la politique apprise
   (h) Un réseau de neurones peut, selon le problème, converger vers la politique optimale;
   VRAi
   (i) Laquelle (une seule) de cette différence temporelle est correcte pour Q :
   1) $Q(s,a) = Q(s,a) + \alpha \cdot (r_t + \max_b Q(s',b) - Q(s,a))$ ← VRAi
   2) $Q(s,a) = Q(s,a) + \alpha \cdot (r_t + Q(s',a) - Q(s,a))$ a tournant dans s' ? faux
   3) $Q(s,a) = Q(s,a) + \alpha \cdot (r_t + \max_b Q(s,b) - Q(s,a))$ le max action dans s tournant n'est pas une diff temporelle
   4) $Q(s,a) = \alpha \cdot (r_t + \max_b Q(s',b) - Q(s,a))$
   La perte des updates passés
   (j) En renforcement, l'agent doit connaitre la fonction de récompense à l'avance;
   FAUX : cf cours

2. Vous êtes conducteur de taxi vers l'aéroport. Votre taxi peut contenir jusqu'à 4 passagers. Chaque passager paye un tarif unique indépedemment de son lieu de prise en charge. L'essence coûte au déplacement, il faut en tenir compte, mais on suppose que notre réservoir à une capacité infinie. L'environnement est totalement observable. Le but est de maximiser son gain.

(a) Quel est l'état terminal ?

l'aéroport

(b) Quelle(s) information(s) mettez-vous dans votre état ?

localisation actuelle + nb de passagers

(c) Comment modélisez-vous votre fonction de récompense ?

- 1 à chaque transition (essence, on veut le + court chemin)
+ prix des courses à l'aéroport

(d) Quelle approche utilisez-vous pour trouver la politique optimale le plus rapidement possible ?

iteration valeur

(e) Désormais, le réservoir n'est plus intarissable, il faut passer à la pompe pour le remplir. Comment modifiez-vous votre état en conséquence ?

etat du plein dans l'etat

(f) Comment modéliser votre fonction de récompense pour prendre en compte le remplissage de la voiture dans votre problème ? Donnez deux changements.

penalité maximum quand plein vide (ie -10000)
penalité faible quand passage à la pompe
(pompe = détour = perte de temps
mais nécessaire)