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;

   (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;

   (c) La fonction Q permet d'apprendre par essai-erreur;

   (d) La Q-learning ne demande aucune connaissance préalable de l'environnement;

   (e) Une instance MDP avec un discount factor γ petit tend à privilégier les récompenses à court terme;

   (f) $V(s)$ est de plus grande dimension que $Q(s, a)$ pour un même problème;

   (g) Pour maximiser sa récompense, il faut progressivement diminuer le facteur d'exploration ϵ au fur et à mesure que la politique s'affine;

   (h) Un réseau de neurones peut, selon le problème, converger vers la politique optimale;

   (i) Laquelle (une seule) de cette différence temporelle est correcte pour Q :

      1) $Q(s, a) = Q(s, a) + \alpha \cdot (r_t + \max_b Q(s', b) - Q(s, a))$

      2) $Q(s, a) = Q(s, a) + \alpha \cdot (r_t + Q(s', a) - Q(s, a))$

      3) $Q(s, a) = Q(s, a) + \alpha \cdot (r_t + \max_b Q(s, b) - Q(s, a))$

      4) $Q(s, a) = \alpha \cdot (r_t + \max_b Q(s', b) - Q(s, a))$

   (j) En renforcement, l'agent doit connaitre la fonction de récompense à l'avance;

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 ?

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

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

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

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

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