Université de Paris

2021-2022

M2 MLDS

Apprentissage par renforcement

Evaluation

Notes :

- Cette évaluation se fera par groupes de 3 à 4 étudiants.
- Pour chaque exercice, vous produirez un document, portant vos noms, et présentant succinctement le problème et le(s) graphique(s) demandés.
- Vous fournirez de plus vos codes
- L'ensemble sera envoyé par mail à l'adresse : marc.metivier@u-paris.fr

Exercice 1 – Prédiction

1. Créer une politique pour Maze appliquant le principe suivant : dans chaque état appliquer les probabilités d'actions {W: 0.1, S:0.4, E:0.4, N:0.1}
2. Comparer MC et TD(0) pour évaluer cette politique dans l'environnement Maze en fonction du nombre d'épisodes effectués. Pour les évaluer, on calculera l'erreur de prédiction par une distance euclidienne entre les vecteurs de valeurs calculés et le vecteur de valeurs produit par IPE. Vous pourrez donc représenter dans un graphiques l'évolution de l'erreur en fonction du nombre d'épisodes effectués.
3. Implémenter n-steps TD(0), puis TD(λ), et les intégrer dans la comparaison (on prendra n=2, 5 et 10)

Exercice 2 – Contrôle

1. Comparer l'évolution du renforcement total par épisode, pour les algorithmes : MC-control, SARSA et Q-Learning dans l'environnement Maze. Vous utiliserez un ε défini de la manière suivante :

$$\varepsilon_t = \begin{cases} \varepsilon_0 \quad si\ t \leq T, \\ \dfrac{\varepsilon_0}{\sqrt{t-T}} \quad sinon. \end{cases}$$

   où ε_0 une valeur initiale et T un nombre d'épisodes pendant lequel ε_t reste constant et égale à ε_0.

   Vous testerez et montrerez les résultats pour différentes valeurs de T et ε_0. Comme résultats vous afficherez des graphiques montrant l'évolution du total des récompenses par épisode et du nombre d'actions exécutées par épisode, au fur-et-à mesure des épisodes.

2. Tester et montrer les performances en utilisant d'autres moyens de gérer l'exploration :
   a. Softmax avec différentes températures
   b. Greedy avec une initialisation optimiste (uniquement pour SARSA et QL)

3. Reproduire ces tests dans les environnements FrozenLake44 et FoorRooms_Key

4. Implémenter UCT et montrer ses performances