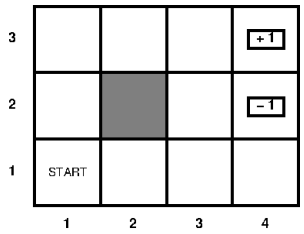
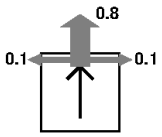
Inteligență artificială
Laborator 8: Reinforcement learning

25 noiembrie 2019

Proces de decizie Markov

- Stări $s \in S$, acțiuni $a \in A$ (starea inițială s_0)
- Modelul de tranziții $P(s'|s, a)$ (probabilitatea de a ajunge din starea s în starea s' efectuând acțiunea a)
- Funcția de recompensă $R(s)$





Învățarea cu întărire (Reinforcement learning, RL) se bazează pe procese de decizie Markov, însă modelul de tranziții și funcția recompensă sunt necunoscute. Algoritmii de tip RL învață o politică optimă. Tipuri de învățare cu întărire:

- învățare pasivă: agentul execută o politică fixă și o evaluează
- învățare activă: agentul își actualizează politica pe măsură ce învață

Învățarea diferențelor temporale (TD) utilizează tranzițiile observate pentru a ajusta utilitățile. Ecuația diferențelor temporale ține cont de diferența utilităților între stări succesive:

$$U^{\pi}(s) \leftarrow U^{\pi}(s) + \alpha(R(s) + \gamma U^{\pi}(s') - U^{\pi}(s))$$

α rata de învățare.

Algoritmul Q-learning învață o funcție acțiune-valoare $Q(a, s)$ (Q *quality*)

- utilitățile $U(s) = max_a Q(a, s)$
- funcțiile Q: un agent TD care învață o funcție Q nu are nevoie de un model de forma $P(s'|s, a)$ pentru învățare sau selecția acțiunii (**model-free**)

Ecuația de actualizare pentru TD Q-Learning:

$$Q(a, s) = Q(a, s) + \alpha(R(s) + \gamma max_{a'} Q(a', s') - Q(a, s))$$

Coeficientul de învățare α determină viteza de actualizare a estimărilor.

Temă

Considerați un labirint care conține obstacole definite. Labirintul poate fi reprezentat printr-o tablă nxn. Un agent se găsește într-un punct inițial. Identificați drumul de la punctul de start până la punctul final

știind că agentul nu poate trece peste obstacole, iar limitele tablei nu pot fi depășite.
Utilizați algoritmul Q-learning pentru a identifica drumul cel mai scurt. Pentru testare considerați un labirint cu cel puțin trei drumuri distincte de la punctul de start pâna la punctul final.

Observație: la fiecare pas agentul trebuie sa decidă care e următoarea mutare. Pentru fiecare acțiune, agentul primește o recompensă (când ajunge la final) sau o penalitate.

Etape
(0.4) 1. Modelarea problemei
(0.6) 2. Implementarea metodei

Bibliografie: Capitolul 21. Reinforcement Learning din AIMA (Artificial Intelligence: A Modern Approach)