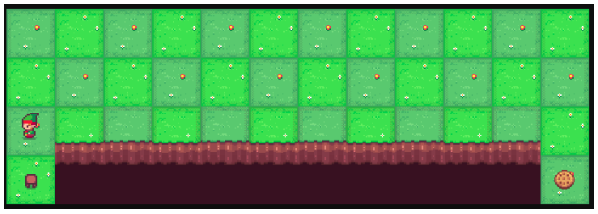
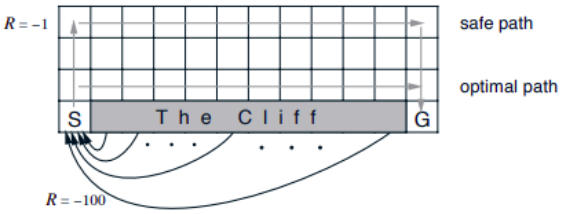
Inteligență Artificială
Laborator 9. Reinforcement Learning

Temă
Considerăm un agent care se poate deplasa într-un mediu (un grid de dimensiuni 4x12). Agentul se poate deplasa în direcția sus, jos, stânga sau dreapta. Punctul de start al agentului este colțul cel mai din stânga jos (3,0). Agentul dorește să ajungă la destinație, în punctul cel mai din dreapta jos (3, 11). Exista o zonă periculoasă (3, 1..10): dacă agentul calcă pe terenul stâncos, atunci se întoarce la punctul de start. Un episod se termină atunci când agentul atinge obiectivul.
Recompensa este -1 pentru toate tranzițiile, cu excepția celor din regiunea maro (3, 1..10) (pentru acestea recompensa este -100).





Implementați algoritmul Q-learning pentru a identifica drumul pe care trebuie să-l parcurgă agentul.
a. (0.1p) inițializarea tabelei Q, a parametrilor algoritmului și a stării inițiale
b. (0.1p) pentru o stare s, identifică starea următoare s' prin aplicarea unei acțiuni a
c. (0.7p) algoritmul Q-learning
- selectează acțiunea cu cea mai mare valoare Q din starea s'
- actualizează valorile Q
- actualizează starea curentă
- repetă
d. (0.1p) afișați politica determinată de algoritm

Bonus: (0.1p) verificați convergența algoritmului (spre ex., un grafic ce conține recompensele în raport cu episodul)

Observație: se poate folosi mediul Open AI gym
https://gymnasium.farama.org/environments/toy_text/cliff_walking/

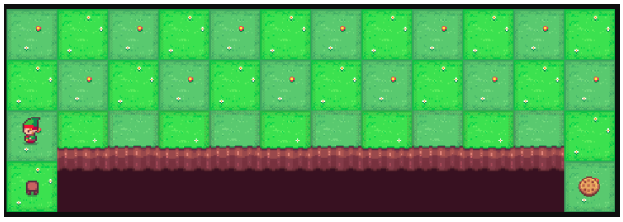
Pentru săptămâna 21-27 noiembrie: punctele a, b
Pentru săptămâna 5-9 decembrie: punctele c,d

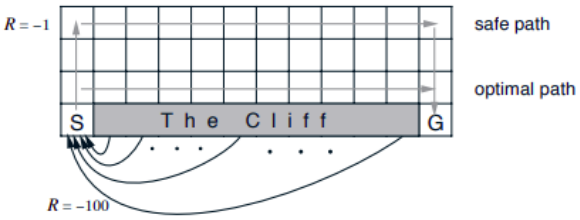
Resurse:
Secțiunea 6.5 Q-learning: Off-policy TD Control
http://incompleteideas.net/book/ebook/node65.html

Homework

Consider an agent that can move in an environment (a 4x12 grid). The agent can move up, down, left or right. The starting point of the agent is the lower leftmost corner (3,0). The agent wants to reach the destination, the bottom rightmost point (3, 11). There is a dangerous area (3, 1..10): if the agent steps on the cliff, it returns to the starting point. An episode ends when the agent reaches the goal.

The reward is -1 for all transitions except those in the region (3, 1..10) (for which the reward is -100).





Implement the Q-learning algorithm to identify the path the agent should take.

a. (0.1p) initialization of Q table, of the algorithm parameters and initial state

b. (0.1p) for a state s, identify the next state s' by applying an action a

c. (0.7p) the Q-learning algorithm

- selects the action with the highest Q value in state s'
- update the Q values
- update the current state
- repeat

d. (0.1p) show the policy determined by the algorithm

Bonus: (0.1p) check the convergence of the algorithm (e.g. a plot containing the rewards over time)

Note: Open AI gym environment can be used
https://gymnasium.farama.org/environments/toy_text/cliff_walking/

For week November 21-27: a, b
For week December 5-9: c,d

Useful Links:
6.5 Q-learning: Off-policy TD Control http://incompleteideas.net/book/ebook/node65.html