Tema 1 - Q-learning

Învățare automată

Nicolăescu Alexandru

Data afișării: 11.03.2018

Data predării: 30.03.2018

Data actualizării: 26.03.2018

Se acceptă maxim 4 zile de întârziere, cu penalizare (din 10 puncte) astfel:

- 0.5p pentru prima zi de întârziere
- 1p/zi pentru următoarele 3 zile

Tema se prezintă la laborator. Limbajul de programare este la alegere.

1 Enunț

Se cere să se implementeze un agent care să joace Pong folosind algoritmul de învățare prin recompensă Q-learning.

Acesta este o variantă 2D simplificată de tenis de masă, în care fiecare din cei doi jucători controlează o paletă pe care o poate mișca sus, jos sau să o lase pe loc. Masa are doi pereți, aflați pe partea superioară și inferioară a acesteia. Atunci când mingea ajunge în același coloană cu o paletă se consideră că jucătorul respectiv pierde. Când agentul câștigă, acesta primește un punct, iar când pierde, i se scade un punct.

Pentru mecanica jocului se poate folosi una dintre variantele următoare:

- Traiectoria mingii se reflectă din pereți și din palete, ea ricoșând simetric, precum în figura 1.
- Logica uzuală de Pong:
  - când mingea lovește paleta în stânga, va ricoșa spre stânga
  - când mingea lovește paleta în centru, va ricoșa perpendicular
  - când mingea lovește paleta în dreapta, va ricoșa spre dreapta

Se va folosi un număr maxim de pași pentru fiecare joc pentru a se evita bucle infinite. Se consideră că este egalitate când se atinge acest prag, scorul rămânând neschimbat.

Mingea se deplasează la unghiuri de 45°, dar nu direct perpendicular către unul din pereți (pentru a se evita o buclă infinită), precum în figura 2. Se consideră că viteza mingii este de o căsuță la fiecare pas, iar la începutul fiecărui punct, mingea este lansată de la centrul mesei într-o direcție aleatoare.

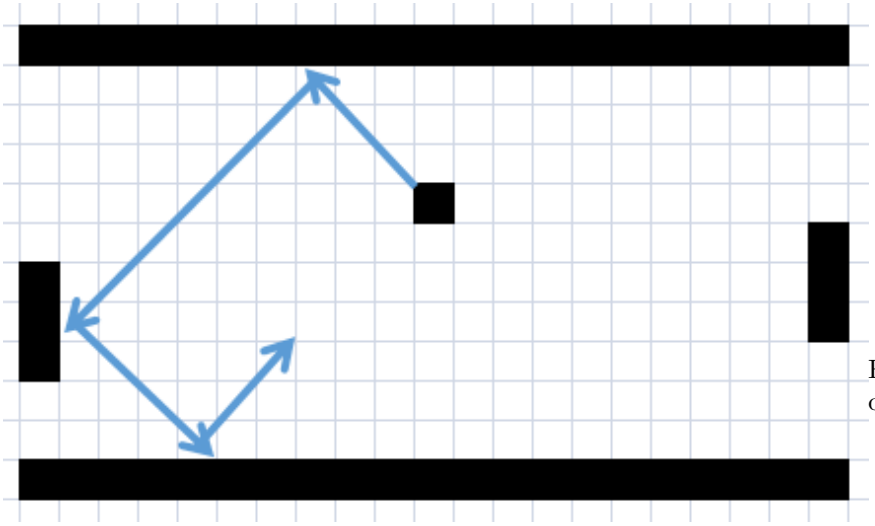

Figura 1: Masa de Pong

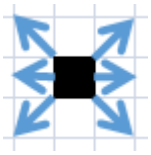

Figura 2: Direcțiile de deplasare ale mingii

2 Cerințe

Se vor testa 3 strategii de explorare/exploatare:

- Strategia aleatoare - agentul alege o acțiune aleatoare la fiecare pas. Deci nu are loc nici o etapă de învățare.
- Strategia lacomă - se alege mereu cea mai bună acțiune.
- Strategia ϵ-greedy - cât timp mai sunt acțiuni neîncercate, se alege aleator din acestea. Altfel, cu probabilitate ϵ se alege aleator, iar cu probabilitate 1-ϵ se alege cea mai bună acțiune.

Se vor folosi următorii adversari:

- Aleator.
- Lacom, generat din agentul curent. În acest caz ambii jucători au același model (aceleași cunoștințe). Agentul construiește Q, ce conține utilitatea fiecărei perechi (stare, acțiune) din punct de vedere al acestuia. Deoarece jocul este simetric, Q poate fi folosit și de adversar pentru a-și alege acțiunea. De exemplu, conceptual, se poate oglindi masa, astfel încât adversarul să fie în locul agentului, iar apoi se determină acțiunea cea mai bună, precum s-ar alege pentru agent.
- Adversar aproape perfect, care cu probabilate α alege acțiunea perfectă, iar cu probabilitate 1 - α alege o acțiune aleatoare. Acțiunea perfectă reprezintă acțiunea unui adversar care joacă ideal. Un exemplu este: deoarece viteza mingii și a paletelor este la fel, de o căsuță la fiecare pas, o euristică simplă este ca paleta adversarului să se miște astfel încât să rămână mereu în dreptul mingii. Astfel, mingea se va lovi mereu de paletă, iar advesarul nu ar putea să piardă.

Se vor folosi aceleași perechi de (Agent, Adversar) atât la etapa de învățare, cât și la cea de testare:

- Agent aleator vs. Adversar aleator.
- Agent lacom vs. Advsersar aleator.
- Agent lacom vs. Adversar lacom - în acest caz jucătorii au același model.

- Agent lacom vs. Adversar aproape perfect
- Agent ϵ-greedy vs. Advsersar aleator
- Agent ϵ-greedy vs. Adversar lacom - în acest caz jucătorii au același model. În etapa de antrenare, agentul alege și acțiuni aleatoare, iar în etapa de evaluare, ambii jucători aleg greedy.
- Agent ϵ-greedy vs. Adversar aproape perfect

Pentru obținearea graficelor, se vor alterna etapele de antrenare și de testare.

Cerințe:

- [**3p**] Generarea sistemului
  - Se cere generarea mesei împreună cu logica de deplasare a mingii, mișcarea paletelor și regulile scorului. Dimensiunea mesei și a paletelor trebuie să fie configurabile.
  - Intefață în mod text sau grafică care să reprezinte desfășurarea acțiunilor.
- [**3p**] Implementare Q-learning
  - [**0.5p**] Strategia aleatoare
  - [**0.5p**] Strategia lacomă
  - [**1p**] Strategia ϵ-greedy
  - [**1p**] Parametrii variabili (rata de învățare, factor de discount)
- [**2.5p**] Grafice
  - [**0.5p**] Influența ratei de învățare
  - [**0.5p**] Influența factorului de discount
  - [**0.5p**] Influența ϵ
  - [**0.5p**] Influența numărului de episoade de antrenare
  - [**0.5p**] Diferența dintre strategii
- [**1.5p**] Discuții
  - Discuție sau grafice asupra influenței adversarului și a configurației mesei: dimensiunea acesteia și a paletelor.
- [**maxim 2p**] bonus
  - [**2p**] Interfață grafică care să prezinte toți parametrii, atât în partea de antrenare, cât și de testare.
  - [**2p**] Implementarea unei metode de grupare pentru a reduce spațiul stărilor.
  - **Punctajele pentru bonus nu se cumulează.**

Graficele și discuțiile se vor regăsi într-un fișier .pdf.

3 Hint-uri

Deoarece spațiul de stări crește exponențial, se recomandă să se înceapă de la un scenariu mic.