Üzleti Intelligencia

4. Előadás: Monte Carlo és temporális különbségek

Kuknyó Dániel
Budapesti Gazdasági Egyetem

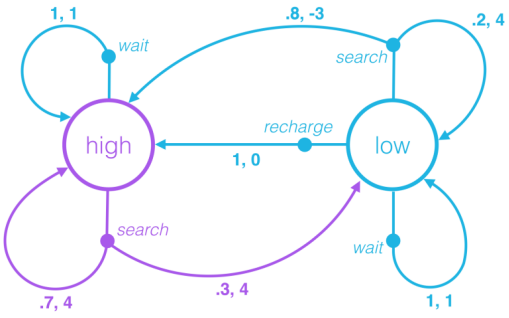
2023/24
1.félév

Az RL modellje

Markov döntési folyamat

$$MDP\,(S, A, P, R, s_0, \gamma)$$

- S: állapotok halmaza
- A: cselekvések halmaza
- $P: S \times A \times S \to [0, 1]$: állapotátmeneti valószínűségek
- $R: S \times A \to \mathbb{R}$: azonnali jutalmak
- s_0: kezdőállapot
- γ: diszkont faktor



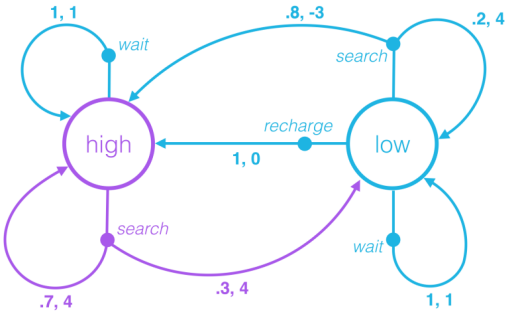
Az RL modellje

Az MDP folyamata:

1. Az ügynök s_0 állapotból indul
2. Az ügynök π politika szerint cselekszik: $a_t \sim \pi(s_t)$
3. A környezet reagál a cselekvésre, és visszaadja az ügynöknek r_{t+1} jutalmat és s_{t+1} következő állapotot
4. Ez ismétlődik amíg a kilépési kritérium be nem teljesül

Cél: Az optimális politika megtalálása. A politika optimális, ha a hozamának várható értéke maximális:

$$E_\pi \left(r_1 + \gamma r_2 + \gamma^2 r_3 + ... \right) \rightarrow max$$



Monte Carlo becslés

A dinamikus programozási algoritmusok futtatásához szükség volt a környezet dinamikájának modelljére. Ez a tulajdonságuk gyakran használhatatlanná teszik őket a gyakorlatban, mert a környezeti dinamika vagy nem ismert vagy nem kiszámítható sok esetben.

A Monte Carlo (**MC**) módszerek ezzel szemben nem igényelnek előzetes tudást a feladat elvégzéséhez: csak **tapasztalat** szükséges ahhoz, hogy megtanuljanak elvégezni egy feladatot. Ezt úgy érik el, hogy mintát vesznek állapotokból, cselekvésekből és jutalmakból, majd az eredményeket átlgolják.

Monte Carlo becslés

Monte Carlo szimuláció

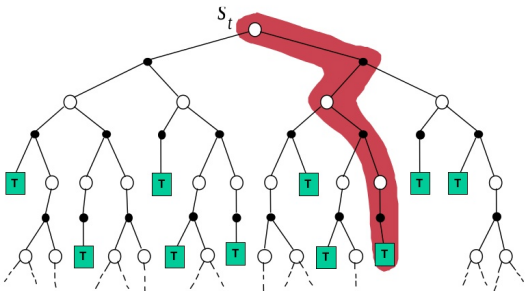
Számítási algoritmusok olyan osztálya, ami a véletlen vagy bizonytalan komponens hatását analizálja vagy szimulálja sztochasztikus folyamatokban.

Monte Carlo a megerősítéses tanulásban

A MC tanítási algoritmus a politika iteráció egy általánosított változata. A megerősítéses tanulásban a MC algoritmusok a mélységi bejárásnak felelnek meg.

A MC módszer egy egyszerű ötletet használ: epizódonkénti nyers tapasztalatok alapján tanul anélkül, hogy modellezné a környezeti dinamikát. A megfigyelt átlagos hozamot a várható megtérülésre tett becslésként számítja ki.

$$V(s_t) \leftarrow V(s_t) + \alpha \left[G_t - V(s_t) \right]$$



Értékfüggvények a Monte Carlo tanításban

A MC mivel epizódonként vesz mintát a hozamokból az értékfüggvények csak sok epizód hozamának átlagolásaként számítódhatnak ki.

Az **állapot-érték függvény** megadja, mennyi a várható hozam, ha az ügynök adott s állapotban áll, és onnan π politikát követi:

$$V_\pi(s) = Avg\left\{G_{t:T} | s_t = s\right\}$$

Ahol

$$G_{t:T} = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

t aktuális időlépéstől a T terminális állapotba vezető időlépésig az ügynök által összegyűjtött diszkontált kumulált hozam.

Értékfüggvények a Monte Carlo tanításban

A MC mivel epizódonként vesz mintát a hozamokból az értékfüggvények csak sok epizód hozamának átlagolásaként számítódhatnak ki.

Az **állapot-cselekvés minőség függvény** megadja, mennyi a várható hozam, ha az ügynök adott s állapotban áll, a cselekvést végrehajtja, majd onnan π politikát követi:

$$Q_\pi(s,a) = Avg\left\{G_{t:T} | s_t = s, a_t = a\right\}$$

Ahol

$$G_{t:T} = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$$

t aktuális időlépéstől a T terminális állapotba vezető időlépésig az ügynök által összegyűjtött diszkontált kumulált hozam.