

Üzleti Intelligencia

Megerősítő tanulás - beadandó feladatok

2023

A feladatok 1-3 skálán vannak osztályozva nehézség szerint, ahol 1 a legkönnyebb és 3 a legnehezebb. A feladatok véletlenszerűen sorsolódnak ki a 6. gyakorlaton. A feladatok beadása Coospace felületen történik, ahol csak egy linket kell beküldeni, ami a feladatot megvalósító Git tárhelyre mutat. Késő beadás nem lehetséges, a dátum beadásakor feltöltött anyagok fognak leosztályozásra kerülni. Minden további információ a tantárgyi útmutatóban és az órán elhangzottakban találhatóak.

1 Duna

Az ügynök Esztergomban áll, és olyan gyorsan kell eljutnia Pestre, amennyire csak lehet. Két út van, amelyekkel elhagyhatja a várost, az 10-es és 11-es út Mindkét út különböző helyre vezet.

Miután megérkezett Budára, két hídon juthat át a Duna folyón, hogy eljusson Pestre. A forgalom kiszámíthatatlan, így előfordulhat, hogy a hídon amit választott forgalmi dugó alakul ki.

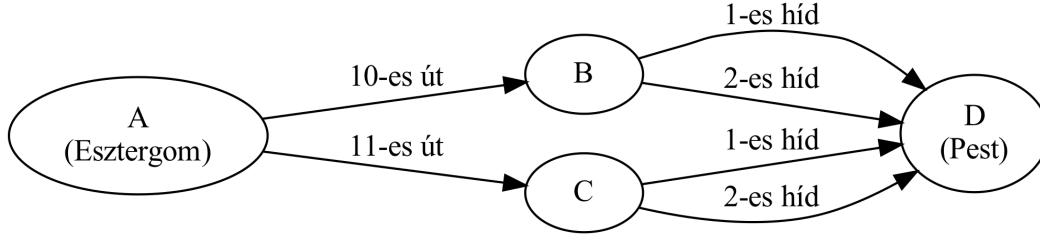
Néha pedig átirányítódik a másik hídra annak ellenére, hogy a másik útvonalat vagy hidat választotta.

A cél az, hogy olyan stratégiát találjon az ügynök, ami a leghamarabb eljuttatja a céljához. A jutalmak a negatív időt jelentik, amelyre szüksége van az adott út/híd átlépéséhez.

Az A , B , C állapotokhoz és minden cselekvéshez a jutalom a következőképpen számolódik ki:

$$R(s, a) = \begin{cases} \text{Ha nincs dugó} & R(s, \text{norm}) \\ \text{Ha dugó van} & R(s, \text{dugó}) \end{cases}$$

Előfordulhat, hogy az ügynök választ egy irányt, de átirányítják a másikra. Tehát ha $P(s, a, s') = 0.9$, tehát 90% valószínűséggel abba az irányba megy, amit választott, de 10% valószínűséggel átirányítják a másik útra.



A környezet dinamikája ($P(s, a, s')$, $R(s, a, s')$):

$$\begin{aligned} P(A, 10, B) &= 0.79, P_{\text{dugó}}(A, 10, B) = 0.9, R_{\text{dugó}}(A, 10, B) = -60, R_{\text{norm}}(A, 10, B) = -21 \\ P(A, 11, C) &= 0.89, P_{\text{dugó}}(A, 11, C) = 0.56, R_{\text{dugó}}(A, 11, C) = -44, R_{\text{norm}}(A, 11, C) = -1 \\ P(B, 1, D) &= 0.84, P_{\text{dugó}}(B, 1, D) = 0.43, R_{\text{dugó}}(B, 1, D) = -20, R_{\text{norm}}(B, 1, D) = -4 \\ P(B, 2, D) &= 0.7, P_{\text{dugó}}(B, 2, D) = 0.84, R_{\text{dugó}}(B, 2, D) = -40, R_{\text{norm}}(B, 2, D) = -11 \\ P(C, 1, D) &= 0.98, P_{\text{dugó}}(C, 1, D) = 0.75, R_{\text{dugó}}(C, 1, D) = -65, R_{\text{norm}}(C, 1, D) = -5 \\ P(C, 2, D) &= 0.77, P_{\text{dugó}}(C, 2, D) = 0.53, R_{\text{dugó}}(C, 2, D) = -58, R_{\text{norm}}(C, 2, D) = -10 \end{aligned}$$

1.1 Dinamikus programozás (2)

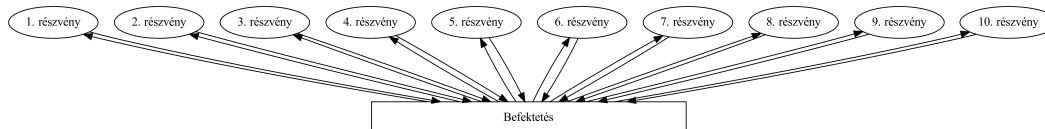
Oldja meg a problémát dinamikus programozással. Ehhez tartozóan implementálja a politika iteráció eljárását (politika kiértékelés, politika javítás). Futtassa a szimulációt adott konvergencia kritériumig. Mutassa meg, hogy az ügynök megtanulta optimalizálni a jutalmat. Írja le röviden az eljárás elméleti alapjait és a saját tapasztalatait egy Jupyter markdown cellában.

1.2 Sztochasztikus becslés (2)

Oldja meg a problémát sztochasztikus becsléssel. Ehhez tartozóan implementálja a Monte Carlo és időbeli különbségek politika keresési eljárásait. Futtassa a szimulációt adott konvergencia kritériumig. Mutassa meg, hogy az ügynök megtanulta optimalizálni a jutalmat. Írja le röviden az eljárás elméleti alapjait és a saját tapasztalatait egy Jupyter markdown cellában.

2 Tőzsde

Az ügynök egy tőzsdén akar befektetni és eladni. A tőzsde úgy működik, hogy minden reggel 10 részvény közül választhat, ami zárás után dől el, hogy jövedelmezett-e vagy nem. Minden nap csak egy részvényt választhat.



Az ügynök a jutalmát a befektetésekből kapja meg. Minden befektetésnek a jutalma egy normális eloszlásból származik, ahol μ a várható érték és σ a variancia.

Részvény	1	2	3	4	5	6	7	8	9	10
μ	0.76	-1.23	1.54	-0.48	1.94	-0.82	-1.27	0.92	-1.99	1.30
σ	0.34	0.72	0.61	0.87	0.28	0.51	0.18	0.43	0.95	0.75

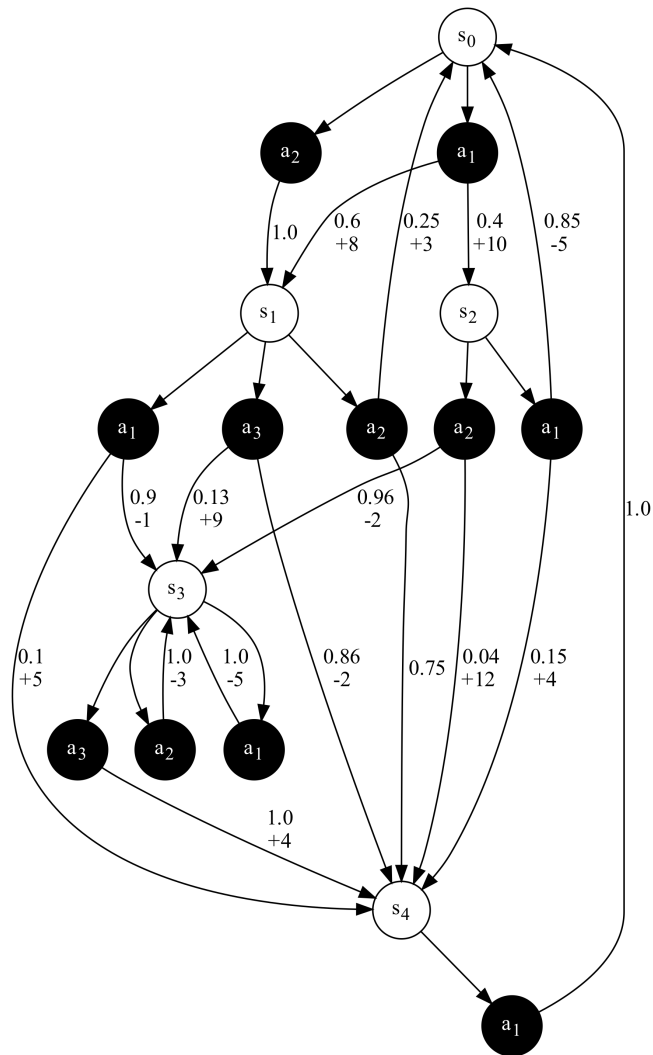
Minden nap az ügynök választhat, hogy a legjobban jövedelmező befektetést választja, vagy kockáztat és véletlenszerűen teszi meg a tétet.

2.1 ε -mohó ügynök (1)

Implementálja a befektetés környezetét és optimalizálja a befektetési stratégia ε értékét. Mutassa meg, hogy az ügynök képes optimalizálni a jutalmat a tanítási iterációk alatt. Adja meg az optimális ε értéket 2 tizedesjegy pontossággal. Hasonlítsa össze a mohó stratégia és az optimális stratégia jutalmait az egyes részvényeken. Ábrázolja az összehasonlítást. Írja le röviden az eljárás elméleti alapjait és a saját tapasztalatait, eredményeit egy Jupyter markdown cellában.

3 MDP

Implementálja az alábbi $MDP(S, A, P, R, s_0, \gamma)$ Markov döntési folyamatot. Az állapotok valószínűsége tizedes törtként szerepel, a jutalmak pedig + és - előjellel a kapcsolatokon. Használja az alábbi paramétereket: $s_0 = 0$, $\gamma = 0.98$



3.1 Bellman egyenletek (1)

Számolja ki a $V(s)$ és $Q(s, a)$ értékeket a Bellman egyenlettel a környezeti dinamikát felhasználva. Mutassa meg, hogy az értékek a valóshoz konvergálnak. Írja le az eljárás elméleti alapjait és a saját tapasztalatait röviden egy Jupyter markdown cellában.

3.2 Q -tanulás (2)

Implementálja a Q -tanulás és dupla Q -tanulás algoritmusát az adott Markov döntési folyamatra. Ábrázolja és hasonlítsa össze a Q -táblákat az értékek konvergenciáját mindkét esetben. Írja le az eljárás elméleti alapjait és a saját tapasztalatait röviden egy Jupyter markdown cellában.