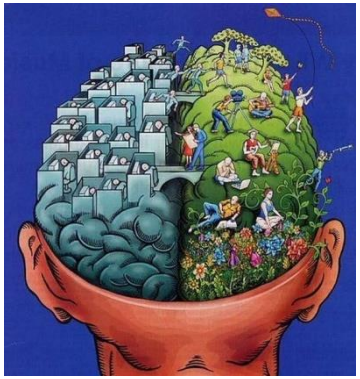
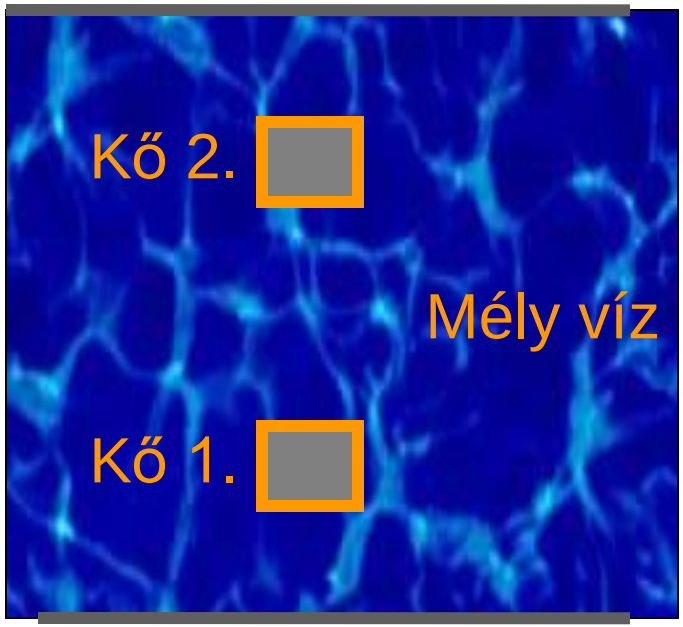
Budapesti Műszaki és Gazdaságtudományi Egyetem
Mesterséges Intelligencia és Rendszertervezés Tanszék



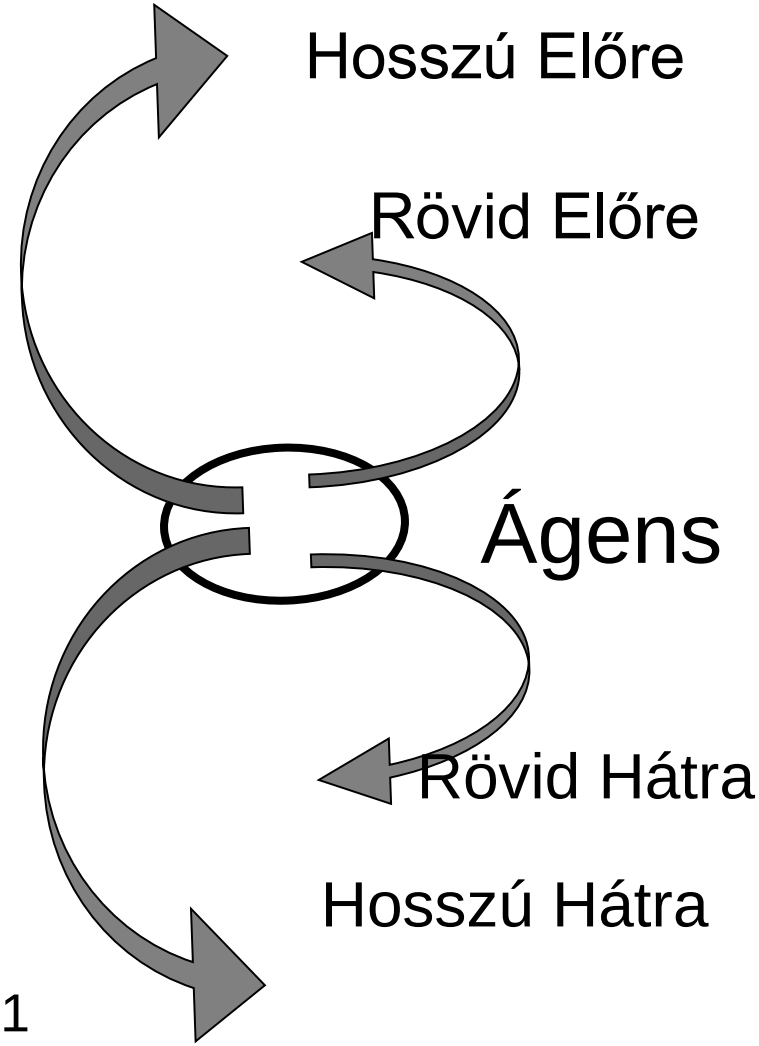
Megerősítéses tanulás - feladatok

Előadó: Dr. Hullám Gábor

Folyópart B



Folyópart A



Tanuló szekvencia:

RE→ HE→ RE→ +1 (száraz lábbal át)

RE→ HE→ HH→ HE→ HH→ RH→ HE→ -1

RE→ HE→ RH→ -1 (megfürdött)

……

Part B

Kő 2.

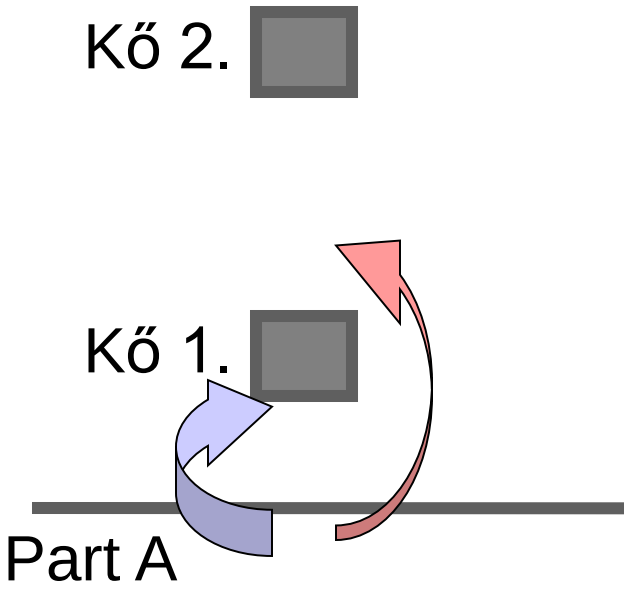
$$Q(a,s) \leftarrow Q(a,s) + \alpha\,(R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s))$$

Kő 1.

Part A

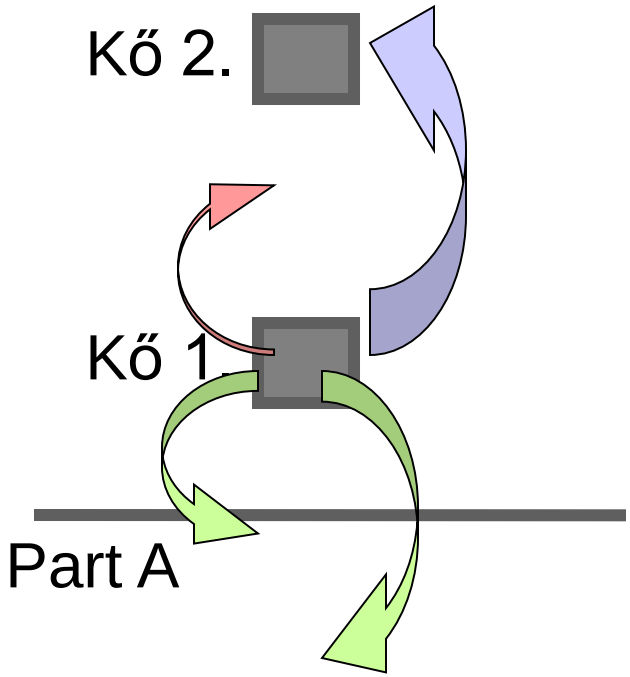
		Cselekvés			
		HH	RH	RE	HE
Állapot	Part A	0	0	-0.035	-0.877
	Kő 1.	-0.520	-0.550	-0.835	0.555
	Kő 2.	-0.204	-0.897	1.180	1.158

Part B

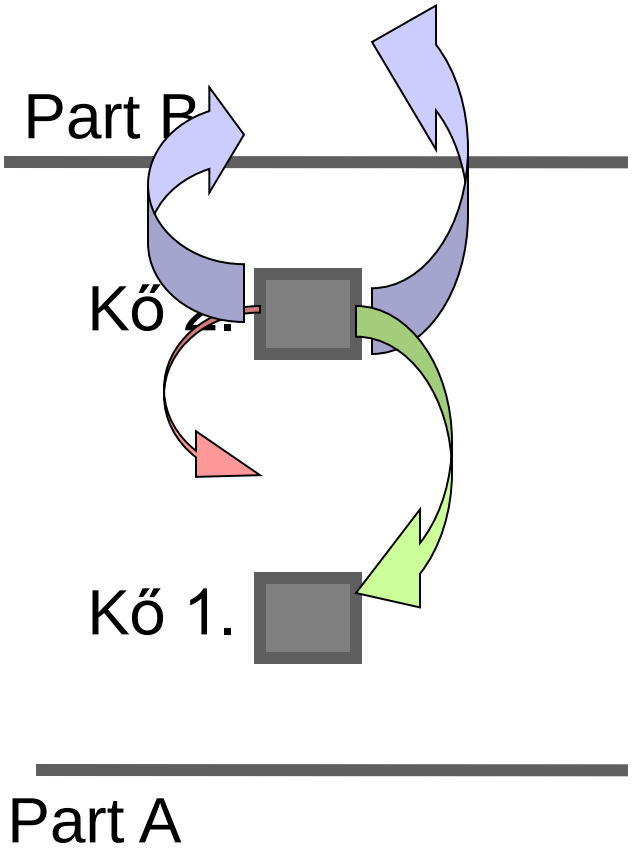


	HH	RH	RE	HE
Part A	0	0	-0.035	-0.877
Kő 1.	-0.520	-0.550	-0.835	0.555
Kő 2.	-0.204	-0.897	1.180	1.158

Part B



	HH	RH	RE	HE
Part A	0	0	-0.035	-0.877
Kő 1.	-0.520	-0.550	-0.835	0.555
Kő 2.	-0.204	-0.897	1.180	1.158



	HH	RH	RE	HE
Part A	0	0	-0.035	-0.877
Kő 1.	-0.520	-0.550	-0.835	0.555
Kő 2.	-0.204	-0.897	1.180	1.158

F5.Legjobb eljárás meghatározása

Aktív megerősítéses tanulást végzünk, a rendelkezésre álló cselekvések *a*1 és *a*2, ezeket az összes
állapotban végrehajthatjuk (kivéve a végállapotot). 8 állapot van (egyszerűen az 1,2,3,…,8 sorszámokkal
jelöltük őket), az állapot-átmenet mátrix az egyes cselekvések választása esetén:

*a*1 választásakor, **P(*s* → *s*'|a1)**

s \ *s*'	1	2	3	4	5	6	7	8
1	0	0,2	0,2	0,1	0	0	0,5	0
2	0	0	0	0	0	0	0	0
3	0,5	0,1	0	0,2	0	0,2	0	0
4	0,1	0	0	0	0,1	0	0	0,8
5	0,5	0	0	0	0	0	0,5	0
6	0	0	0	0,2	0,2	0	0,2	0,4
7	0	0	0	0,3	0,2	0,5	0	0
8	0	0	0,6	0	0,2	0,2	0	0

*a*2 választásakor, **P(*s* → *s*'|a2)**

s \ *s*'	1	2	3	4	5	6	7	8
1	0	0,2	0	0,1	0	0,2	0,3	0,2
2	0	0	0	0	0	0	0	0
3	0,1	0,4	0	0,5	0	0	0	0
4	0,1	0	0,6	0	0,1	0	0	0,2
5	0,5	0,1	0	0	0,2	0	0,2	0
6	0	0	0	0,2	0,4	0	0	0,4
7	0,1	0	0	0,3	0,2	0,2	0,2	0
8	0	0	0,4	0	0,2	0,2	0	0,2

Az állapotok valódi hasznosságértékei:

s	1	2	3	4	5	6	7	8
U(s)	+1	+20	-4	-10	0	+10	+3	5

Adja meg az adott állapotra vonatkozó optimális eljárásmódot meghatározó képletet!
Mi lesz az optimális eljárásmódunk (optimális stratégiánk) az **s**= 3-al jelölt állapotban?

F5.Legjobb eljárás meghatározása

*a*1 választásakor, **P(*s* → *s*'|a1)**

s \ *s*'	1	2	3	4	5	6	7	8
1	0	0,2	0,2	0,1	0	0	0,5	0
2	0	0	0	0	0	0	0	0
3	0,5	0,1	0	0,2	0	0,2	0	0
4	0,1	0	0	0	0,1	0	0	0,8
5	0,5	0	0	0	0	0	0,5	0
6	0	0	0	0,2	0,2	0	0,2	0,4
7	0	0	0	0,3	0,2	0,5	0	0
8	0	0	0,6	0	0,2	0,2	0	0

*a*2 választásakor, **P(*s* → *s*'|a2)**

s \ *s*'	1	2	3	4	5	6	7	8
1	0	0,2	0	0,1	0	0,2	0,3	0,2
2	0	0	0	0	0	0	0	0
3	0,1	0,4	0	0,5	0	0	0	0
4	0,1	0	0,6	0	0,1	0	0	0,2
5	0,5	0,1	0	0	0,2	0	0,2	0
6	0	0	0	0,2	0,4	0	0	0,4
7	0,1	0	0	0,3	0,2	0,2	0,2	0
8	0	0	0,4	0	0,2	0,2	0	0,2

Az állapotok valódi hasznosságértékei:

s	1	2	3	4	5	6	7	8
U(s)	+1	+20	-4	-10	0	+10	+3	5

F5.Legjobb eljárás meghatározása

*a*1 választásakor, **P(*s* → *s*'|a1)**

s \ *s*'	1	2	3	4	5	6	7	8
1	0	0,2	0,2	0,1	0	0	0,5	0
2	0	0	0	0	0	0	0	0
3	0,5	0,1	0	0,2	0	0,2	0	0
4	0,1	0	0	0	0,1	0	0	0,8
5	0,5	0	0	0	0	0	0,5	0
6	0	0	0	0,2	0,2	0	0,2	0,4
7	0	0	0	0,3	0,2	0,5	0	0
8	0	0	0,6	0	0,2	0,2	0	0

*a*2 választásakor, **P(*s* → *s*'|a2)**

s \ *s*'	1	2	3	4	5	6	7	8
1	0	0,2	0	0,1	0	0,2	0,3	0,2
2	0	0	0	0	0	0	0	0
3	0,1	0,4	0	0,5	0	0	0	0
4	0,1	0	0,6	0	0,1	0	0	0,2
5	0,5	0,1	0	0	0,2	0	0,2	0
6	0	0	0	0,2	0,4	0	0	0,4
7	0,1	0	0	0,3	0,2	0,2	0,2	0
8	0	0	0,4	0	0,2	0,2	0	0,2

Az állapotok valódi hasznosságértékei:

s	1	2	3	4	5	6	7	8
U(s)	+1	+20	-4	-10	0	+10	+3	5

$$U(s) = R(s) + \gamma \max_a \sum_{s'} T(s,a,s') \cdot U(s')$$

$$0{,}5 \cdot U(s_1) + 0{,}1 \cdot U(s_2) + a_2 \cdot U(s_4) + 0{,}2 \cdot U(s_6)$$

$$0{,}5 + 2 + -2 + 2 = 2{,}5$$

$$0{,}1 \cdot U(s_1) + 0{,}4 \cdot U(s_2) + 0{,}5 \cdot U(s_4)$$

$$0{,}1 + 8 - 5 = 3{,}1$$

$$11 * < a_2$$

F6.Q-tanulás

Egy robot Q-tanulással tanulja az optimális eljárásmódot. A robot környezete 2 db S1 és S2 állapotból áll.
Mindegyik állapotban 2 db a1 és a2 cselekvést lehet alkalmazni. A tanulási tényező (bátorsági faktor) és a leszámoltatási tényező egyformán 1/2. A robot 4 db példát dolgoz fel:

I. (kiindulás = S1, cselekvés = a1, jutalom = 10, vége = S2)
II. (kiindulás = S2, cselekvés = a2, jutalom = -10, vége = S1)
III. (kiindulás = S1, cselekvés = a2, jutalom = 10, vége = S1)
IV. (kiindulás = S1, cselekvés = a1, jutalom = 10, vége = S1)

Frissítse fel futamonként a Q-érték táblázatát (a táblázat eredetileg 0-ra legyen inicializálva). Adja meg az alkalmazott frissítési egyenletet!

$Q(a,s) \leftarrow Q(a,s) + \alpha \left[R(s) + \gamma \max_{a'} Q(a',s') - Q(a,s) \right]$

$s = s_1, \; s' = s_2 \quad a = a_1 \quad R = 10$

$Q(a_1,s_1) = 0$

$\frac{1}{2} \cdot 10$

$Q(a^*,s_2) = 0$

$S = 0 + \frac{1}{2}(10 + 0 + 0)$

Kiindulás:

	a1	a2
S1	0	0
S2	0	0

I. példa:

	a1	a2
S1	?	?
S2	?	?

F6.Q-tanulás

Egy robot Q-tanulással tanulja az optimális eljárásmódot. A robot környezete 2 db S1 és S2 állapotból áll.
Mindegyik állapotban 2 db a1 és a2 cselekvést lehet alkalmazni. A tanulási tényező (bátorsági faktor) és a
leszámoltatási tényező egyformán 1/2. A robot 4 db példát dolgoz fel:
I. (kiindulás = S1, cselekvés = a1, jutalom = 10, vége = S2)
II. (kiindulás = S2, cselekvés = a2, jutalom = -10, vége = S1)
III. (kiindulás = S1, cselekvés = a2, jutalom = 10, vége = S1)
IV. (kiindulás = S1, cselekvés = a1, jutalom = 10, vége = S1)
Frissítse fel futamonként a Q-érték táblázatát (a táblázat eredetileg 0-ra legyen inicializálva). Adja meg az
alkalmazott frissítési egyenletet!

Kiindulás:

	a1	a2
S1	0	0
S2	0	0

I. példa:

	a1	a2
S1	?	?
S2	?	?

$s = s_2 \quad a = a_2 \quad s' = s_1 \quad R(s) = -10$

$Q(a_2, s_2) = \emptyset$

$Q(a_2, s_2) = Q(a_2, s_2) + \alpha\left(R(s_2) + \gamma \max_{a'} Q(a', s_1) - Q(a_2, s_2)\right)$

$= 0 + \frac{1}{2}\left(-10 + \left(\frac{1}{2} \cdot 5\right) - 0\right) = -3{,}25$

F6.Q-tanulás

Egy robot Q-tanulással tanulja az optimális eljárásmódot. A robot környezete 2 db S1 és S2 állapotból áll.
Mindegyik állapotban 2 db a1 és a2 cselekvést lehet alkalmazni. A tanulási tényező (bátorsági faktor) és a
leszámoltatási tényező egyformán 1/2. A robot 4 db példát dolgoz fel:
I. (kiindulás = S1, cselekvés = a1, jutalom = 10, vége = S2)
II. (kiindulás = S2, cselekvés = a2, jutalom = -10, vége = S1)
III. (kiindulás = S1, cselekvés = a2, jutalom = 10, vége = S1)
IV. (kiindulás = S1, cselekvés = a1, jutalom = 10, vége = S1)
Frissítse fel futamonként a Q-érték táblázatát (a táblázat eredetileg 0-ra legyen inicializálva). Adja meg az
alkalmazott frissítési egyenletet!

Kiindulás:

	a1	a2
S1	0	0
S2	0	0

I. példa:

	a1	a2
S1	?	?
S2	?	?

$S = S_1 \quad S' = S_1 \quad a = a_2 \quad R = 10$

$$Q(a_2, S_1) = \underbrace{Q(a_2 S_1)}_{\emptyset} + \alpha \left[R(s_1) + \gamma \max_{a'} Q(a', S') - \underbrace{Q(a_2 S_1)}_{\emptyset} \right]$$

$$10 + \frac{1}{2} \cdot 5$$

$$= \emptyset + \frac{1}{2} \left[10 + \frac{1}{2} \cdot 5 - \emptyset \right] = 6{,}25$$

④ $s = s_1$ $s = s_1$ $a = a_1$ $R(s_1) = 10$

$Q(a_1, s_1) = 5$

$$Q(a_1, s_1) = \underbrace{Q(a_1, s_1)}_{5} + \alpha \left[R(s) + \gamma \max_{a'} Q(a', s_1) - Q(a_1, s_1) \right]$$

$$5 + \frac{1}{2}\left[10 + \left(\frac{1}{2} \cdot 6{,}25 \right) - 5 \right]$$

$$= 9{,}0625$$