Mesterséges Intelligencia MI

Szekvenciális döntési probléma

(Megerősítéses tanulás)

Pataki Béla

BME I.E. 414, 463-26-79

pataki@mit.bme.hu,

http://www.mit.bme.hu/general/staff/pataki

Ismétlés – félév eleji előadások.

Tanulás alapvető fajtái:

felügyelt tanulás egy esetnél mind a bemenetet, mind a kimenetet észlelni tudjuk (bemeneti minta + kívánt válasz)

- döntési fák
- neurális hálók

megerősítéses tanulás az ágens az általa végrehajtott tevékenység csak bizonyos értékelését kapja meg, esetleg nem is minden lépésben (jutalom, büntetés, **megerősítés**)

felügyelet nélküli tanulás semmilyen információ sem áll rendelkezésünkre a helyes kimenetről (az észlelések közötti összefüggések tanulása)

féligellenőrzött tanulás a tanításra használt esetek egy részénél mind a bemenetet, mind a kimenetet észlelni tudjuk (bemeneti minta + kívánt válasz), a másik – tipikusan nagyobb – részénél csak a bemeneti leírás ismert

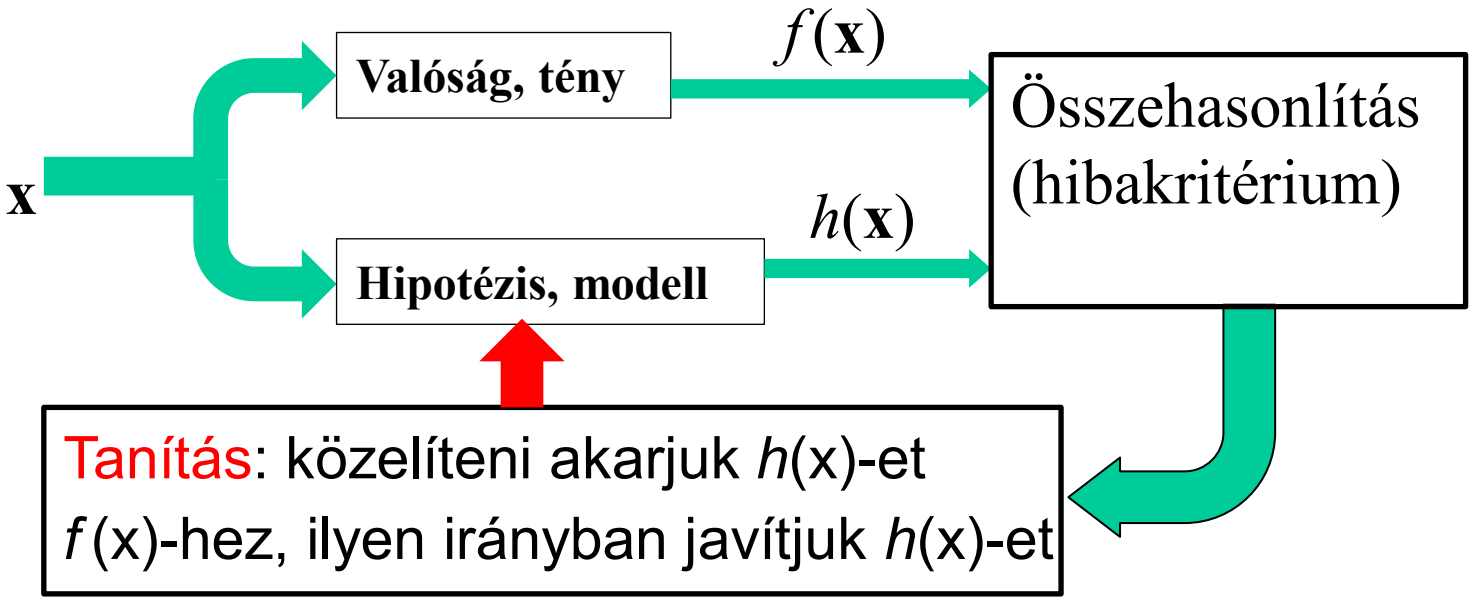
Felügyelt tanulás (pl. induktív következtetés):

tanulási példa: $(\mathbf{x}, f(\mathbf{x}))$ adatpár, ahol $f(\mathbf{x})$ ismeretlen

tanulás célja: $f(\mathbf{x})$ értelmes közelítése egy $h(\mathbf{x})$ hipotézissel

$h(\mathbf{x}) = f(\mathbf{x})$, $\mathbf{x}$ – ismert példákon gyakran teljes pontosság

$h(\mathbf{x}') \cong f(\mathbf{x}')$, **$\mathbf{x}'$ – a tanulás közben még nem látott esetek (általánosító képesség)**



Feladat: *az ismeretlen **f**-re vonatkozó példák egy halmaza alapján (**tanítóhalmaz**), adjon meg egy olyan **h** függvényt (**hipotézist**), amely tulajdonságaiban jól közelíti az **f**-et (amit **teszthalmazon** verifikálunk).*

Eddigiekben egy-egy mintához tartozott egy-egy jó válasz, kívánt eredmény. Az $f(\mathbf{x})$ alapvetően statikus volt: a pillanatnyi bemenethez keresünk jó választ az adott pillanatban.

Fontos problémacsoport, amikor egy ***lépéssorozat*** mentén ***csak időnként kapunk visszajelzést***, hogy jó vagy rossz, amit csinálunk.

- Triviális példa a sakk, csak a végén derül ki pontosan, hogy jó volt-e a lépéssorozatunk.
- Amikor az egyetemről hazamegyünk, egy sor döntést hozunk, hogy milyen járművel, merre menjünk – a végén derül ki, hogy gyorsan vagy lassan értünk-e haza.

A lépéssorozatot állapotsorozatként modellezzük, pl:

S0 $\Rightarrow$ S103 $\Rightarrow \bullet \quad \bullet \quad \bullet \Rightarrow$ S53 $\Rightarrow \bullet \bullet \bullet \Rightarrow$ S17 $\Rightarrow \bullet \bullet \bullet \Rightarrow$ S68

Bizonyos lépéseknél kapunk jutalmat (**R**eward, ez lehet negatív jutalom, büntetés is), pl.

R(17)= **+2**, R(103)=-17, R(53**)=+100** a többinél nincs

A lépéssorozat hasznossága (additív jut.) $\Rightarrow$+2-17+100=+85

Szekvenciális döntési probléma

Kezdőállapot: **s0** **Kvíz**

Állapotátmenet-modell: **T(*s*, a, *s′*):** ***s*→*s′*** átmenet valószínűsége, amikor az ***a*** cselekvést választottuk s-ben

Jutalom: **R(*s*)** (Nálunk most: additív!)

Mivel az *s* állapotból bizonyos valószínűséggel jutunk különböző lépéssorozatokon át végállapotba, ezért nem tudjuk megmondani, hogy mennyi lesz most a pontos összjutalom a végállapotig. Csak a várható értékét (várható jutalom átlagosan) tudjuk kiszámítani.

Például kétféle sorozat indulhat ki *s*-ből, az egyik valószínűsége 0,3 a másiké 0,7. Az első sorozatban -24 „jutalmat” kapunk, míg a végállapotba érünk, a másik sorozatban +73 jutalmat.

$$VárhatóJutalom(s) = 0,7 \cdot 73 - 0,3 \cdot 24 = 43,9$$

Kvíz

Egy orvosnál megjelenik Ló Leopold paciens, aki erősen köhög.
A dokinak 3 lehetősége van: felír neki rózsaszín, zöld vagy fekete pirulát, abban a reményben, hogy a paciens meg fog gyógyulni.
Az orvosi példánkban mit jelent modellünk valószínűségi jellege?

A. A paciens csak bizonyos valószínűséggel veszi be az előírt pirulát
B. A paciens nem veszi be az előírt pirulát, de bizonyos valószínűséggel meggyógyul
C. Az előírt pirula csak bizonyos valószínűséggel gyógyítja meg a pacienst
D. A paciens bizonyos valószínűséggel másik pirulát vesz be

Szekvenciális döntési probléma

Az s állapot hasznossága: **U(s) a várható hátralévő jutalom**, ami az innen kiinduló lépéssorozatokban (súlyozott) átlagosan elérhető. (Nem az adott állapotban, hanem a végállapotig tartó lépéssorozat során összegezve!)

Az elérhető jutalom függ attól, hogy milyen cselekvéseket választunk!

Az s állapot* U(s) *hasznossága – optimális cselekvések esetén!

Optimális eljárásmód: $a(s) = \pi^*(s)$ megadja minden állapotra, hogy melyik az a cselekvés az adott állapotban, ami maximalizálja a hátralévő jutalmat.

Példa:

Kezdőállapot: s11
Végállapotok: s6 és s8 – csak itt van jutalom, másutt 0.

Az s11, s9, s7, s16, s10 állapotokban választható a ’**Fel**’ cselekvés.

Az s1, s2, s3, s4, s5, s11, s12, s13, s14, s15 állapotokban a ’**Jobbra**’ cselekvést választhatjuk.

Az s11 kezdőállapotban csak bizonyos valószínűséggel teljesül a választott cselekvés, a többi állapotban 100%-ban az történik, amit akarunk (’Fel’, illetve ’Jobbra’).

T(s11,a=’Fel’,s9)=0,8 T(s11,a=’Fel’,s12)=0,2
T(s11,a=’Jobbra’,s9)=0,1 T(s11,a=’Jobbra’,s12)=0,9

s1	s2	s3	s4	s5	s6, R(6)=+10
s7	X	X	X	X	s8, R(8)=-4
s9	X	X	X	X	s10
s11	s12	s13	s14	s15	s16

Mekkora lesz s11 hasznossága – U(11)?

s1	s2	s3	s4	s5	s6, R(6)=+10
s7	X	X	X	X	s8, R(8)=-4
s9	X	X	X	X	s10
s11	s12	s13	s14	s15	s16

Tegyük fel, hogy a ’**Fel**’ cselekvést választjuk s11-ben,

- ha jó irányba indultunk, utána s6-ba navigálhatjuk magunkat,
- ha s12-be jutottunk, akkor nem tudunk mit csinálni, előbb-utóbb s8-ba jutunk.

A várható nyereség ez esetben: 0,8*10-0.2*4=**+7,2**

Ha a ’**Jobbra**’ cselekvést választjuk s11-ben,

- ha mégis felfele indultunk, utána s6-ba navigálhatjuk magunkat,
- ha s12-be jutottunk, akkor nem tudunk mit csinálni, előbb-utóbb s8-ba jutunk.

A várható nyereség ez esetben: 0,1*10-0.9*4= **-2,6**

Rajtunk áll, hogy milyen cselekvést választunk s11-ben: a várható hátralévő összjutalom (az s11 hasznossága) **U(11)=+7,2**

Fontos egyszerűsítés: a sorozat hasznossága = a sorozat állapotaihoz rendelt hasznosságok összege. (a hasznosság **additív)**

Az állapot **hátralevő-jutalma** (**reward-to-go**) az adott lépéssorozatban: azon jutalmak összege, amelyet akkor kapunk, ha az adott állapotból valamelyik végállapotig eljutunk.

Egy állapot várható hasznossága = a hátralevő-jutalom várható értéke

$$U^{\pi}(s) = E\left\{\sum_k R(s_k) \,|\, \pi, s_0 = s\right\}$$

(közönségesen: várható érték=az előfordulási gyakorisággal súlyozott átlag)

Függ attól, hogy milyen az eljárásmódunk (milyen cselekvéseket választunk az egyes állapotokban) – ezt fejezi ki a π felső index ($a(s)=\pi(s)$) .

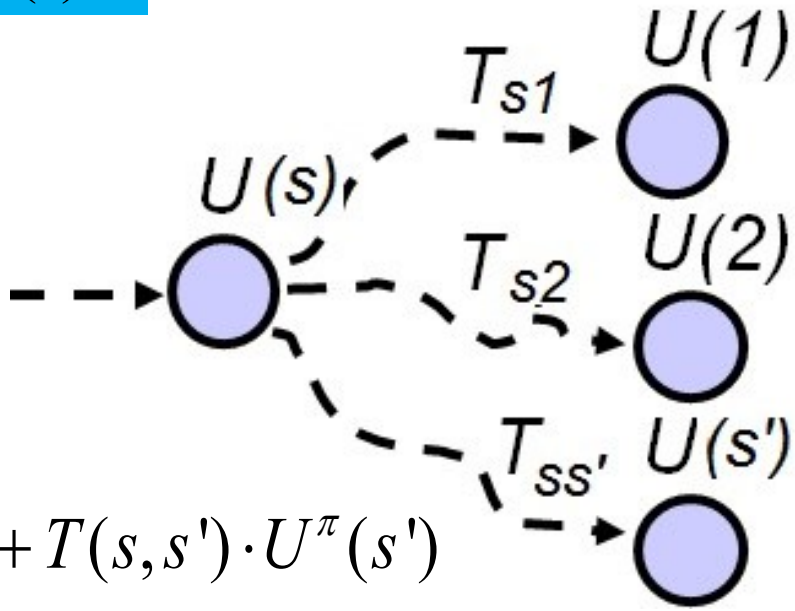
Viszont az s-ből kiinduló út mindig két részre bontható: s-ből a következő állapotig, majd a következőből a végállapotig

$$s0 \Rightarrow s1 \Rightarrow s2 \Rightarrow s3 \Rightarrow \ldots\ldots \Rightarrow sVég$$

ebből számoljuk U(s1)-t

ebből számoljuk U(s)-t

Kapcsolat van az aktuális és az aktuálist követő állapotok hasznossága közt!



$$U^{\pi}(s) = R(s) + T(s,s_1) \cdot U^{\pi}(s_1) + \ldots + T(s,s') \cdot U^{\pi}(s')$$

Szekvenciális döntési probléma

Markov döntési folyamat (MDF) a ϵ A – lehetséges cselekvések

s0- kiinduló (start) állapot, s ϵ S – állapottér elemei

T(s, a, s′) –áll. átmenet valószínűség

R(s) – jutalomfüggvény

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

γ - leszámítolási tényező („jobb ma egy veréb, mint holnap egy túzok”)

Optimális eljárásmód $$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U(s')$$

Szekvenciális döntési probléma

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

A valós $U(s)$-t akkor tudjuk kiszámítani, ha az optimális eljárásmódot ismerjük, DE az optimális eljárásmódot csak akkor tudjuk meghatározni, ha $U(s)$-t ismerjük! Nemlineáris (mindegyik egyenletben van „max” !) egyenletrendszert kéne megoldani!

Értékiteráció (t változó az iterációs lépések számlálója):

1. $t = 0$ - valamilyen kiinduló hasznosságfüggvény $U_0(s)$
2. $U_{t+1}(s)$ meghatározása (nem oldjuk meg az egyenletrendszert, csak kiszámítunk mindegyikből egy-egy új $U(s)$ értéket, a max-ot persze használjuk!)

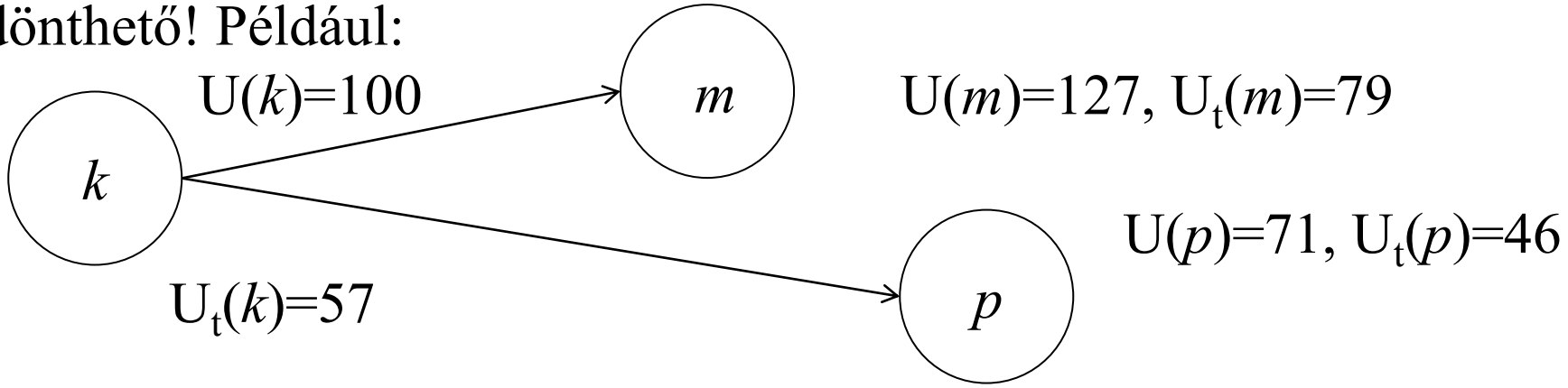
$$U_{t+1}(s) \leftarrow R(s) + \gamma \max_a \sum_{s'} T(s,a,s') \cdot U_t(s')$$

3. $t \leftarrow t+1$
4. Ha már egyik $U(s)$-nél sincs változás (vagy egy adott értéknél kisebb) - KÉSZ , ha volt változás, folytassuk újra a 2.-nél

Konvergens eljárás!

Eljárásmód-iteráció:

A hasznosságértékek lassan konvergálnak az értékiterációnál, de a *legjobb cselekvés* rendszerint már jóval a pontos konvergencia előtt eldönthető! Például:



Már látható a *t*-dik lépésben is, hogy a k→m átmenetet eredményező cselekvések a jobbak, pedig az $U_t(s)$ értékek még messze vannak a valóstól!

A *t*-dik iterációs lépésben az eljárásmódunk $\pi_t(s)$ rögzített (a jelenleg ismert hasznosságok alapján). Ekkor az $U_t(s)$-hez nem kell a nemlineáris max, mivel rögzítettük az eljárásmóddal, hogy jelenleg (*t*-dik iteráció) milyen cselekvést választunk. Tehát *n* állapot esetén egy *n* egyenletből álló lineáris egyenletrendszert kell csak megoldanunk!

$$U_t(s) = R(s) + \gamma \sum_{s'} T(s, \pi_t(s), s') \cdot U_t(s')$$

Eljárásmód-iteráció:

Ötlet: a *t*-dik iterációs lépésben az eljárásmódunk $\pi_t(s)$ rögzíthetjük (a jelenleg ismert hasznosságok alapján).

Ekkor az $U_t(s)$-hez nem kell a nemlineáris max, mivel rögzítettük az eljárásmóddal, hogy jelenleg (*t*-dik iteráció) milyen cselekvést választunk. Tehát *n* állapot esetén egy *n* egyenletből álló lineáris egyenletrendszert kell csak megoldanunk!

$$U_t(s) = R(s) + \gamma \sum_{s'} T(s, \pi_t(s), s') \cdot U_t(s')$$

$$U_t(s_1) = R(s_1) + \gamma \sum_{k=1}^{n} T(s_1, \pi_t(s_1), s_k) \cdot U_t(s_k)$$

$$U_t(s_2) = R(s_2) + \gamma \sum_{k=1}^{n} T(s_2, \pi_t(s_2), s_k) \cdot U_t(s_k)$$

$$\bullet \quad \bullet \quad \bullet$$

$$U_t(s_n) = R(s_n) + \gamma \sum_{k=1}^{n} T(s_n, \pi_t(s_n), s_k) \cdot U_t(s_k)$$

Szekvenciális döntési probléma

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

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U(s')$$

Eljárásmód-iteráció:

1. $t = 0$ - valamilyen kiinduló eljárásmód $\pi_0(s)$
2. $U_t(s)$ meghatározása az $U_t(s) = R(s) + \gamma \sum_{s'} T(s,\pi_t(s),s') \cdot U_t(s')$ lineáris egyenletrendszerből (n állapot, n egyenlet)
3. Amelyik s-re $\max_a \sum_{s'} T(s,a,s') \cdot U_t(s') > \sum_{s'} T(s,\pi_t(s),s') \cdot U_t(s')$

   arra $\pi_{t+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \cdot U_t(s')$ a többire $\pi_{t+1}(s) = \pi_t(s)$
4. $t \leftarrow t+1$
5. Ha már egyik s-nél sincs változás - KÉSZ , ha volt változás, akkor folytassuk újra a 2.-nél