Introducere în Reinforcement Learning

Cursul #2

Ștefan Iordache, Cătălina Iordache, Ciprian Păduraru

Cuprins

01 Concepte generale

Primii pași în RL

02 Procese Markov

Baza algoritmilor noștri!

03 Algoritmi RL

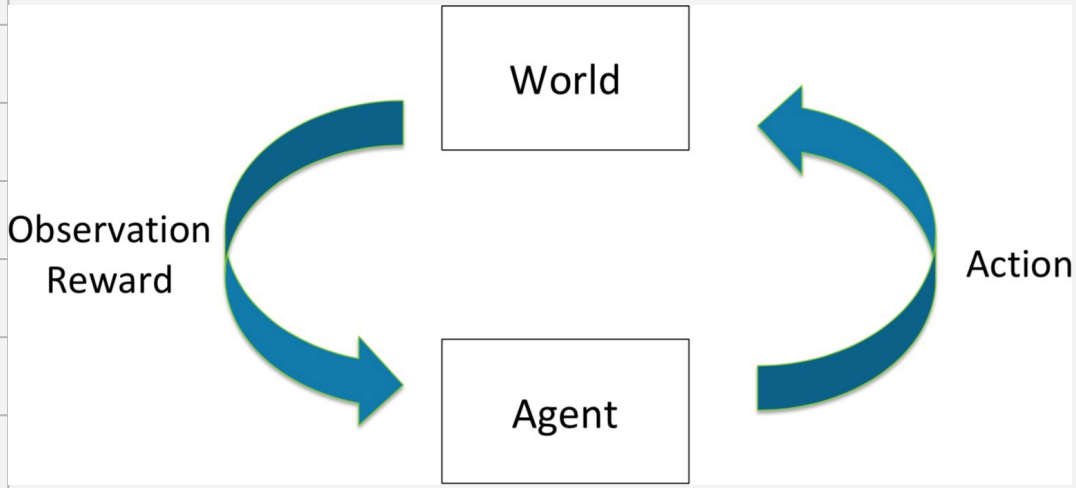
Începe distracția!

01

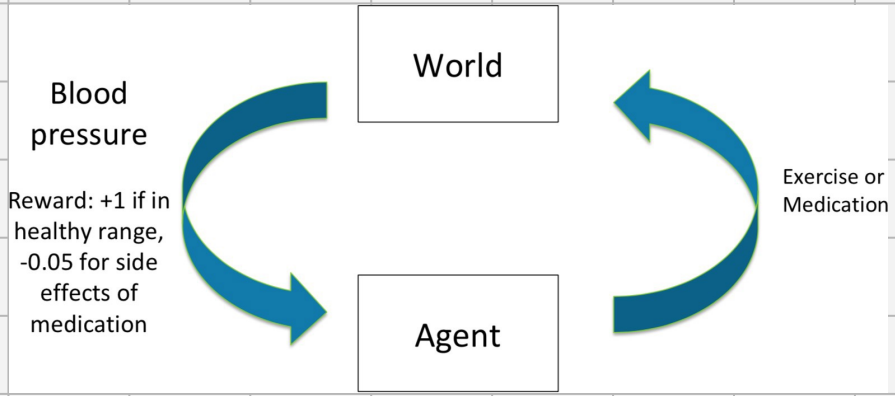
Concepte generale

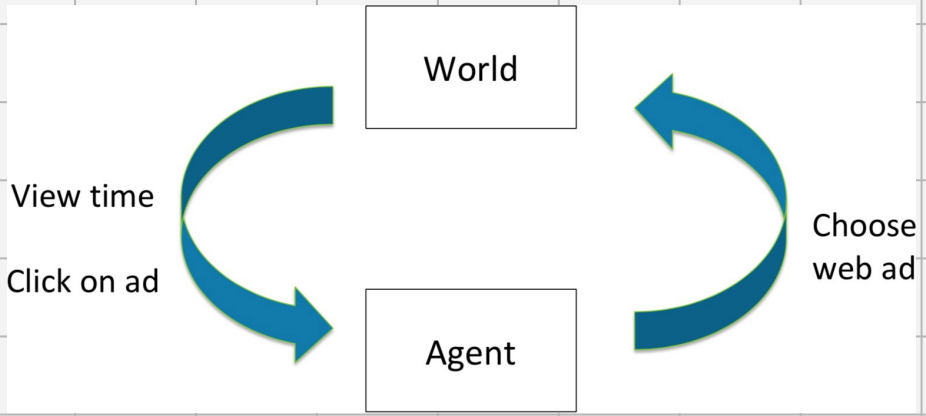
Primii pași în RL

Diagrama generică - RL

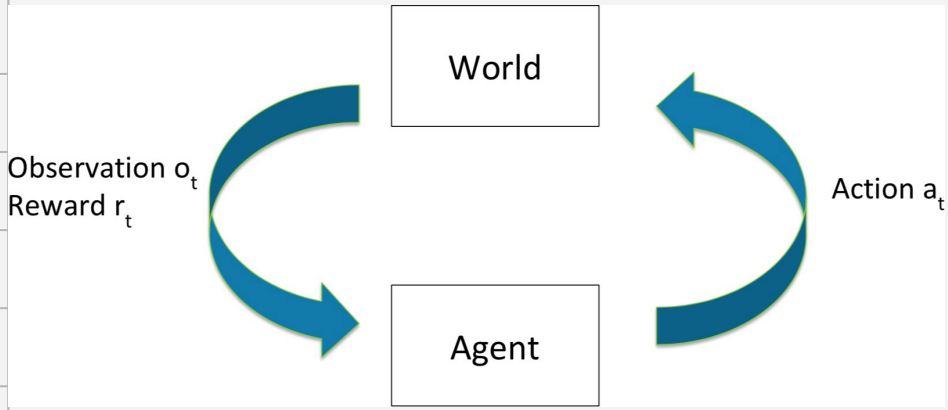


- **Scopul**: alegerea acțiunilor care *măresc recompensele viitoare*.
- Necesită ***balansarea*** recompenselor pe *termen scurt și lung*.


World
Blood pressure
Reward: +1 if in healthy range, -0.05 for side effects of medication
Exercise or Medication
Agent


World
View time
Click on ad
Choose web ad
Agent

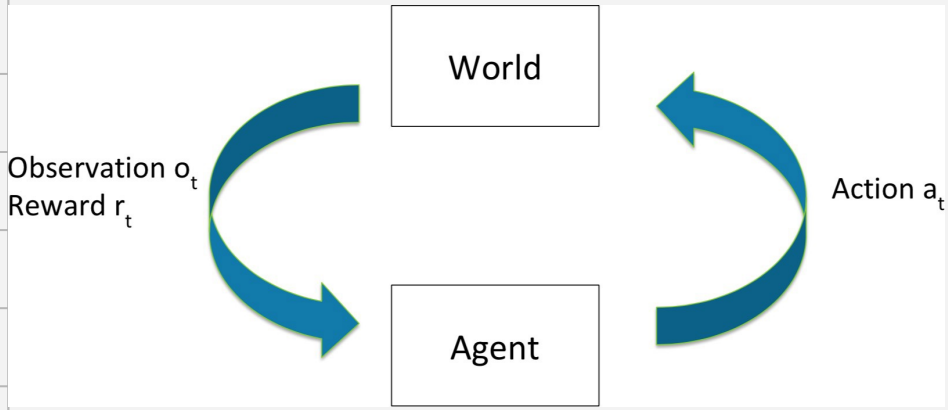
Procesul decizional: Agentul & Mediul



Pentru fiecare moment **t** de timp:

- Agentul execută o **acțiune a_t**
- Mediul este actualizat în urma acțiunii a_t și emite **observația o_t**, respectiv **recompensa r_t**
- Agentul **primește** cele două rezultate din mediu: observația și recompensa

Istoricul observațiilor



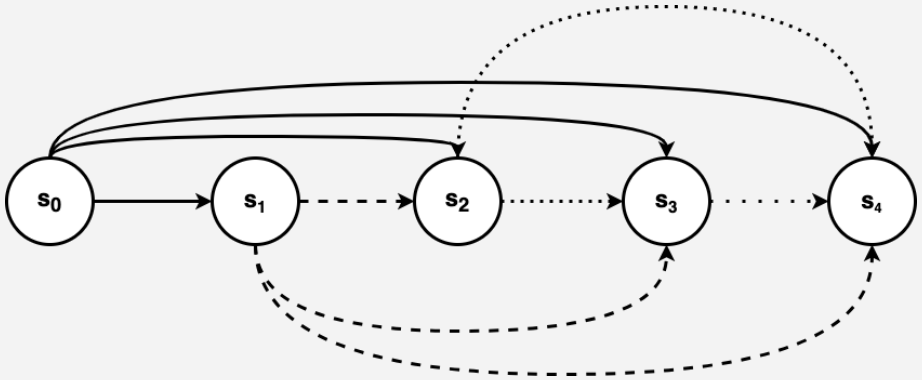
Istoricul la un moment de timp determinat, t:

$$h_t = \{a_1, o_1, r_1, \dots, a_t, o_t, r_t\}$$

- **Acțiunile sunt alese în baza istoricului. Cum?**

În baza unei funcții $\mathbf{s_t = f(h_t)}$

Procese Stochastice



Dinamica unui astfel de proces:

$$Pr(s_t | s_{t-1}, s_{t-2}, \dots, s_0)$$

- **Setul de stări este notat cu S.**
- **Problema: orizont infinit al funcției Pr.**

Care este soluția???

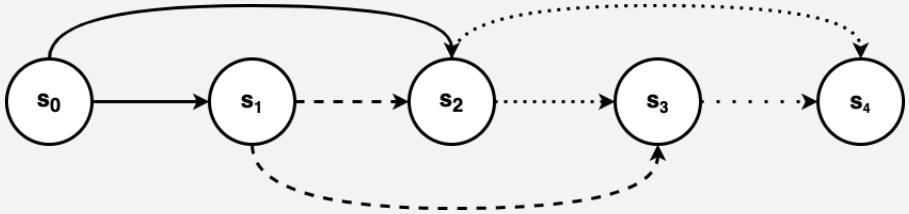
- **Procese Markov**

02

Procese Markov

Baza algoritmilor
noştri!

Procese Markov



Starea curentă depinde doar de un set finit de stări din trecut.

- **Procese Markov de ordinul I**

$$Pr(s_t|\ s_{t-1}, \dots, s_0) = Pr(s_t|\ s_{t-1})$$

- **Procese Markov de ordinul II**

$$Pr(s_t|\ s_{t-1}, \dots, s_0) = Pr(s_t|\ s_{t-1}, s_{t-2})$$

Procese Markov – staționaritate

În mod implicit, un proces Markov se referă la cel de ordinul I.

$$Pr(s_t|\, s_{t-1}, \dots, s_0) = Pr(s_t|\, s_{t-1})\ \forall t$$

- Extindem reprezentarea de mai sus către termenul de ***proces staționar:***

$$Pr(s_t|\, s_{t-1}) = Pr(s_{t'}|\, s_{t'-1})\ \forall t'$$

Avantajul unui proces staționar:

reprezentarea simplă! $Pr(s'|s)$

Întrebări naturale...

- **Cum reprezentăm stările? Cât de multe caracteristici adăugăm?**
  - Răspuns: Până când procesul poate fi considerat Markovian, respectiv staționar.

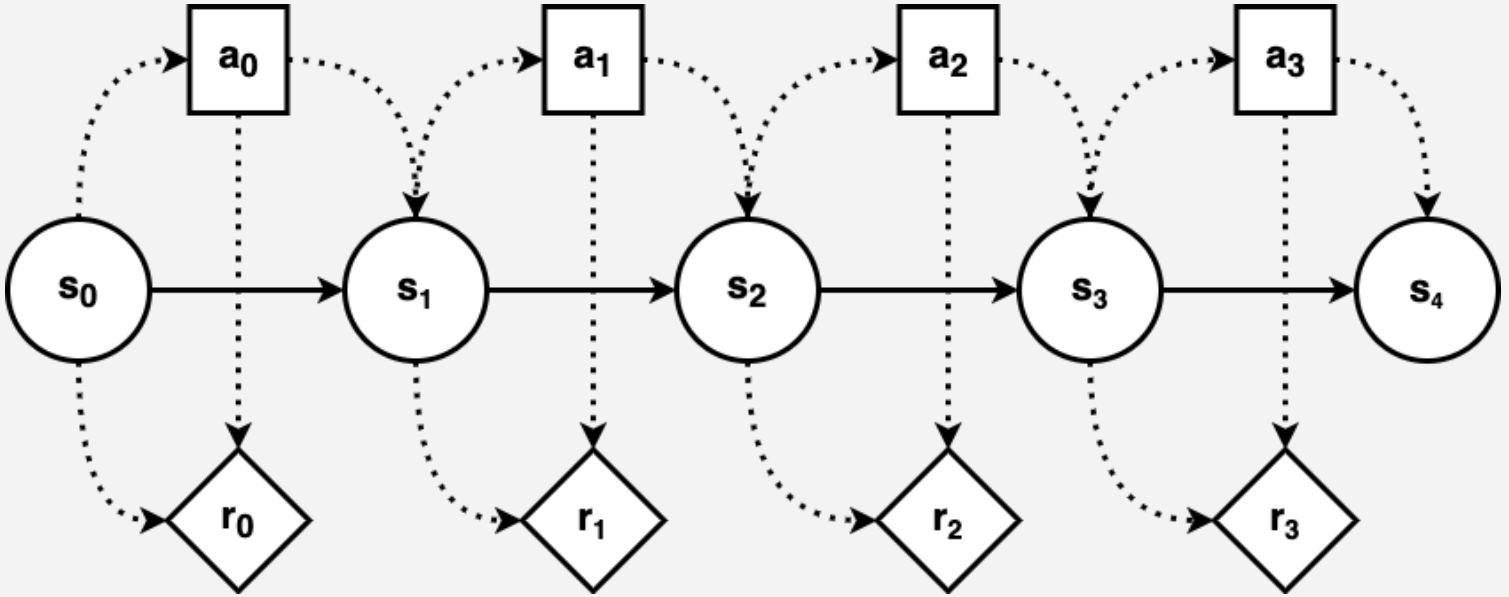
- **Există posibilitatea să adăugăm prea multe componente unei stări?**
  - Răspuns: Da! Adăugarea de componente va crește complexitatea calculelor, implicit necesarul computațional.
  - Soluția: Căutam cel mai mic subset de caracteristici care descrie procesul complet procesul Markovian!

- **Ce utilizăm în practică?**
  - Răspuns: Cel mai frecvent vom utiliza următoarea presupunere -> $s_t = o_t$

Despre decizii

- **Considerăm faptul că simplele predicții sunt inutile. De ce???**
  - Răspuns: Dorim să obținem informații care vor influența alegerile viitoare, nu simple predicții utile unui singur moment.
- **Astfel, sarcina noastră constă în conceperea unor algoritmi capabili să furnizeze decizii!**
- **Dar, cum influențăm deciziile? Cum construim acest algoritm?**
  - Răspuns: ***Procese Decizionale Markov!***

Procese Decizionale Markov



Întrebări naturale... (partea 2)

- **Stările sunt în continuare de tip Markov?**
- **Este lumea parțial observabilă?**
- **Dinamica este deterministă sau stochastică?**
- **Acțiunile influențează recompensa imediată sau afectează recompensele și starea următoare?**

03

Algoritmi RL

Începe distracția!

Algoritmi RL

Cuprind una sau mai multe dintre următoarele componente:

- **Model**
- **Politică (Policy)**
- **Value Function**

Modelul

- **Este reprezentarea agentului pentru felul în care mediul se va schimba în urma unei anumite acțiuni.**

- **Tranzițiile (felul în care agentul prezice următoarea stare):**

$$p(s_{t+1} = s' | s_t = s, a_t = a)$$

- **Metodologia de predicție a recompenselor:**

$$r(s_t = s, a_t = a) = E[r_t | s_t = s, a_t = a]$$

Politica (Policy) - π

- **O politică determină felul în care agentul alege acțiunile pe care le execută. Este o funcție de forma**

$$\pi: S \rightarrow A$$

- **Politici deterministe:** $\pi(s) = a$
- **Politici stochastice:** $\pi(a|s) = Pr(a_t = a|s_t = s)$

Value Function– V^{π}

- **Reprezintă suma recompenselor (cu discount), sub o anumită politică aplicată.**

$$V^{\pi}(s_t = s) = E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \cdots | s_t = s]$$

- **Factorul de discount (γ) va stabili importanța recompensei imediate și a celor viitoare.**
- **Metoda poate fi utilizată pentru a decide calitatea anumitor stări și acțiuni, ulterior stabilind o metodă de comparație între diverse politici.**

Thanks!

Este timpul pentru întrebări!!!

stefan.iordache10@s.unibuc.ro
catalina.patilea@s.unibuc.ro
ciprian.paduraru@fmi.unibuc.ro

+40 7.. ... ...