

Sistemi Intelligenti Reinforcement Learning: Equazioni di Bellman

Alberto Borghese

Università degli Studi di Milano
Laboratorio di Sistemi Intelligenti Applicati (AIS-Lab)
Dipartimento di Scienze dell'Informazione
borghese@dsi.unimi.it



A.A. 2010-2011

1/35

<http://homes.dsi.unimi.it/~borghese/>



Sommario



Determinazione della value function. Esempio.

Le equazioni di Bellman

A.A. 2010-2011

2/35

<http://homes.dsi.unimi.it/~borghese/>



Reinforcement Learning Problem



Given: Repeatedly...

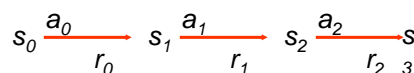
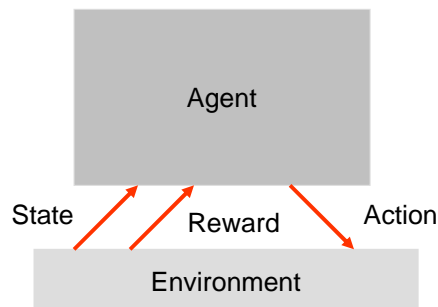
- Executed action
- Observed state
- Observed reward

Learn action policy $\pi: S \rightarrow A$

- ◆ Maximizes life reward $r_0 + \gamma r_1 + \gamma^2 r_2 \dots$ from any start state.
- ◆ Discount: $0 < \gamma < 1$

Note:

- Unsupervised learning
- Delayed reward



Goal: **Learn** to choose actions that maximize life reward

$$r_0 + \gamma r_1 + \gamma^2 r_2 \dots$$



Il nostro filo logico



La Value function ci serve per decidere l'azione migliore.
Per calcolare la Value function devo collezionare reward futuro.

Come se ne esce?

Determinazione algebrica della Value Function

Determinazione "esplorativa" della Value Function

Determinazione della Value Function e scelta della policy via via sempre più intrecciate tra loro.



Esempio: AIBO search



Azioni:

- 1) Rimanere fermo e aspettare che qualcuno getti nel cestino una lattina vuota.
- 2) Muoversi attivamente in cerca di lattine.
- 3) Tornare alla sua base (recharge station) e ricaricarsi.

Stato:

- 1) Alto livello di energia.
- 2) Basso livello di energia.

Goal: collezionare il maggior numero di lattine.

Policy:

$A(s = \text{high}) = \{\text{Search}, \text{Wait}\}$

$A(s = \text{low}) = \{\text{Search}, \text{Wait}, \text{Recharge}\}$

A.A. 2010-2011

5/35

<http://homes.dsi.unimi.it/~borghese/>



Funzionamento del Robot



Funzione Stato prossimo:

$$P_{s \rightarrow s' | a} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$

Se il livello di energia è alto ($s_t = \text{alto}$):

se scelgo Wait - $s_{t+1} = \text{alto}$.

se scelgo Search, s_{t+1} avrà una certa probabilità di diventare low.

$$P_{\text{high} \rightarrow \text{low} | \text{Search}} = \Pr\{s_{t+1} = \text{low} | s_t = \text{high}, a_t = \text{Search}\} = \alpha$$

Se il livello di energia è basso ($s_t = \text{basso}$):

se scelgo Wait - $s_{t+1} = \text{basso}$.

se scelgo Recharge - $s_{t+1} = \text{alto}$.

se scelgo Search, s_{t+1} avrà una certa probabilità di fermarsi.

$$P_{\text{low} \rightarrow \text{low} | \text{Search}} = \Pr\{s_{t+1} = \text{low} | s_t = \text{low}, a_t = \text{Search}\} = \beta$$

A.A. 2010-2011

6/35

<http://homes.dsi.unimi.it/~borghese/>



Reward del Robot



Funzione Reward:

$$R_{s \rightarrow s'|a} = E\{r_{t+1} = r^i | s_t = s, a_t = a, s_{t+1} = s'\}$$

R^{search} reward se il robot sta cercando.

R^{wait} reward se il robot sta cercando.

R^{die} se occorre portarlo a ricaricarsi.

R^{recharge} se il robot va autonomamente a ricaricarsi.

$$R^{\text{search}} > R^{\text{wait}} > R^{\text{recharge}} > R^{\text{die}}$$

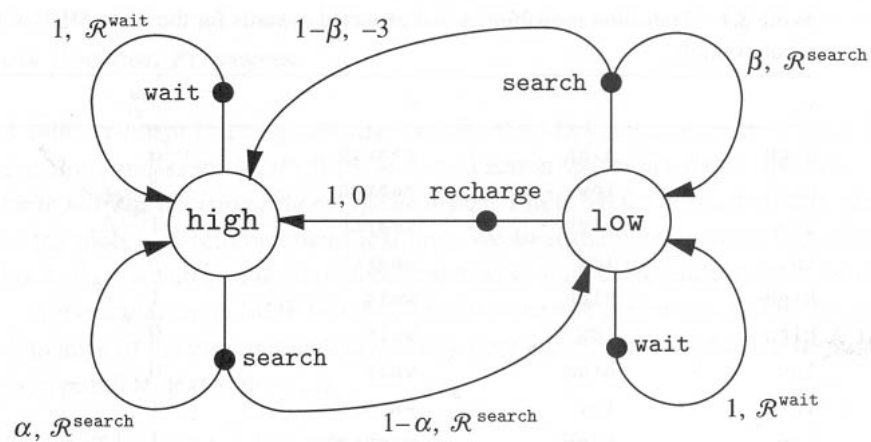
A.A. 2010-2011

7/35

<http://homes.dsi.unimi.it/~borghese/>



State Transition Graph



Probabilistic Finite State Machine

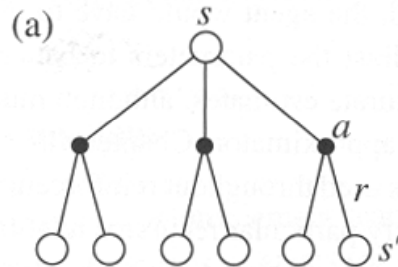
A.A. 2010-2011

8/35

<http://homes.dsi.unimi.it/~borghese/>



Policy



La policy deve essere ancora determinata. Come fa l'agente a determinare la policy ottimale?

Archi multipli fuoriuscenti da un'azione sono associati alla probabilità di scegliere quel cammino (ambiente stocastico).

Archi multipli fuoriuscenti da uno stato, sono associati alla policy.



Value function & policy

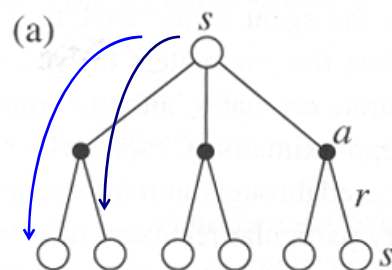
Nulla è detto sulla policy: dato uno stato, in quale nodo azione mi sposto?

Vogliamo costruire agenti lungimirar

State-Value function
(function of the policy):

$$V^\pi(s) = E_\pi \{R_t \mid s_t = s\} =$$

$$E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$



Massimizzo la ricompensa a lungo termine, $V(\cdot)$. Dipende dalla policy:



Value function e modelli markoviani

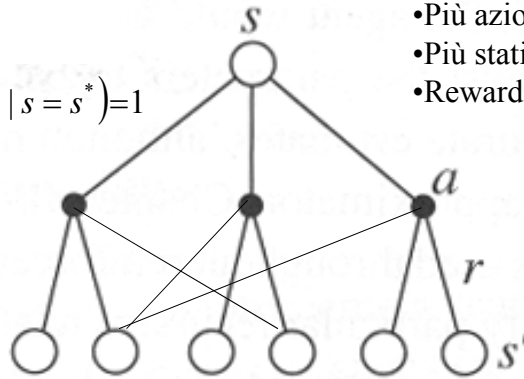


Anche la policy può essere stocastica.

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.

$$\sum_{j=1}^{N_{\text{azioni}}} \Pr(a_j | s = s^*) = 1$$



$$\sum_{k=1}^{N_{\text{stati}}} \Pr(s_{t+1} = s_k | s_t = s'; a_t = a_j) = 1$$

$$V^\pi(s) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

A.A. 2010-2011

11/35

<http://homes.dsi.unimi.it/~borghese/>

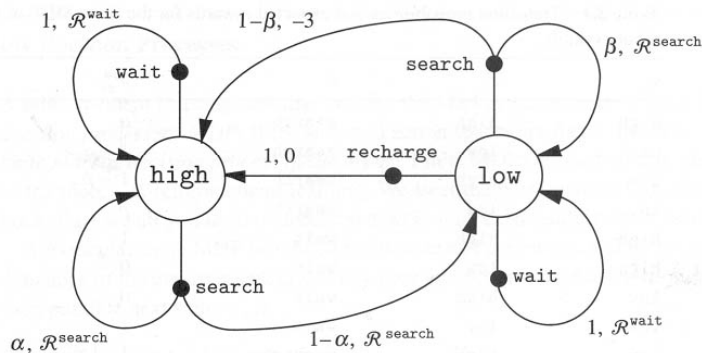


Esempio di calcolo della Value function



Value function
 $V(\text{high}) = ?$
 $V(\text{low}) = ?$

Policy
 $a(\text{high}) = ?$
 $a(\text{low}) = ?$



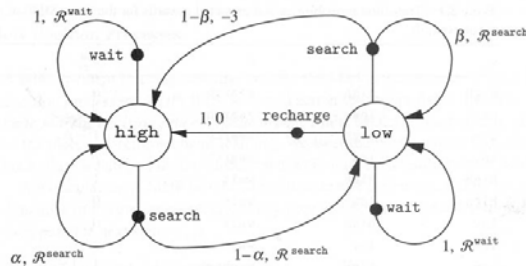
A.A. 2010-2011

12/35

<http://homes.dsi.unimi.it/~borghese/>



Esempio di calcolo della funzione valore



$\alpha=0.4, \beta=0.1, \gamma=0.8, R_{search}=3, R_{wait}=1$

$$\begin{aligned}
 V_h &= \underbrace{\Pr(W)}_{Wait} \times 1 \times [1 + 0.8V_h] + \underbrace{\Pr(S)}_{Search \rightarrow high} \times 0.4 \times [3 + 0.8V_h] + \underbrace{\Pr(S)}_{Search \rightarrow low} \times 0.6 \times [3 + 0.8V_l] \\
 V_l &= \underbrace{\Pr(W)}_{Wait} \times 1 \times [1 + 0.8V_l] + \underbrace{\Pr(S)}_{Search \rightarrow Low} \times 0.1 \times [3 + 0.8V_l] + \underbrace{\Pr(S)}_{Search \rightarrow High} \times 0.9 \times [-3 + 0.8V_h] + \\
 &\quad \underbrace{\Pr(R)}_{Recharge} \times 1 \times [0 + 0.8V_h]
 \end{aligned}$$

Sistema lineare di 2 equazioni nelle 2 incognite: V_h e V_l
Come calcolo V_h e V_l ?

A.A. 2010-2011

13/35

<http://homes.dsi.unimi.it/~borghese/>



Esempio di calcolo della funzione valore - I



$$\begin{aligned}
 V_h &= \underbrace{\Pr(W)}_{Wait} \times 1 \times [1 + 0.8V_h] + \underbrace{\Pr(S)}_{Search \rightarrow high} \times 0.4 \times [3 + 0.8V_h] + \underbrace{\Pr(S)}_{Search \rightarrow low} \times 0.6 \times [3 + 0.8V_l] \\
 V_l &= \underbrace{\Pr(W)}_{Wait} \times 1 \times [1 + 0.8V_l] + \underbrace{\Pr(S)}_{Search \rightarrow Low} \times 0.1 \times [3 + 0.8V_l] + \underbrace{\Pr(S)}_{Search \rightarrow High} \times 0.9 \times [-3 + 0.8V_h] + \\
 &\quad \underbrace{\Pr(R)}_{Recharge} \times 1 \times [0 + 0.8V_h]
 \end{aligned}$$

Devo specificare una policy (stocastica):

$s = \text{high}$ [$\Pr(W) = 0.4$ $\Pr(S) = 0.6$]

$s = \text{low}$ [$\Pr(W) = 0.4$ $\Pr(S) = 0.5$ $\Pr(R) = 0.1$]

$$\begin{aligned}
 V_h &= 0.4 \times 1 \times [1 + 0.8V_h] + 0.6 \times 0.4 \times [3 + 0.8V_h] + 0.6 \times 0.6 \times [3 + 0.8V_l] \\
 V_l &= 0.4 \times 1 \times [1 + 0.8V_l] + 0.5 \times 0.1 \times [3 + 0.8V_l] + 0.5 \times 0.9 \times [-3 + 0.8V_h] + \\
 &\quad 0.1 \times 1 \times [0 + 0.8V_h]
 \end{aligned}$$

A.A. 2010-2011

14/35

<http://homes.dsi.unimi.it/~borghese/>



Esempio di calcolo della funzione valore - II



Devo specificare una policy (stocastica):

$$s = \text{high} \quad [\text{Pr}(W) = 0.4 \quad \text{Pr}(S) = 0.6]$$

$$s = \text{low} \quad [\text{Pr}(W) = 0.4 \quad \text{Pr}(S) = 0.5 \quad \text{Pr}(R) = 0.1]$$

$$V_h = 0.4 \times 1 \times [1 + 0.8V_h] + 0.6 \times 0.4 \times [3 + 0.8V_h] + 0.6 \times 0.6 \times [3 + 0.8V_l]$$

$$V_l = 0.4 \times 1 \times [1 + 0.8V_l] + 0.5 \times 0.1 \times [3 + 0.8V_l] + 0.5 \times 0.9 \times [-3 + 0.8V_h] + 0.1 \times 1 \times [0 + 0.8V_h]$$

$$0.488 V_h = 2.20 + 0.288 V_l \quad \Rightarrow \quad V_h \approx 6,35$$

$$0.64 V_l = -0.8 + 0.44 V_h \quad \Rightarrow \quad V_l \approx 3,12$$

Cosa si può concludere? $V_h > V_l$. Ma non molto altro.



Esempio di calcolo della funzione valore - III



Stessa policy (stocastica) ma altro ambiente (robot che si scarica più difficilmente):

$$s = \text{high} \quad [\text{Pr}(W) = 0.4 \quad \text{Pr}(S) = 0.6]$$

$$s = \text{low} \quad [\text{Pr}(W) = 0.4 \quad \text{Pr}(S) = 0.5 \quad \text{Pr}(R) = 0.1]$$

$$\alpha = 0.4, \beta = 0.9, \gamma = 0.8, R_{\text{search}} = 3, R_{\text{wait}} = 1$$

$$V_h = 0.4 \times 1 \times [1 + 0.8V_h] + 0.6 \times 0.4 \times [3 + 0.8V_h] + 0.6 \times 0.6 \times [3 + 0.8V_l]$$

$$V_l = 0.4 \times 1 \times [1 + 0.8V_l] + 0.5 \times 0.9 \times [3 + 0.8V_l] + 0.5 \times 0.1 \times [-3 + 0.8V_h] + 0.1 \times 1 \times [0 + 0.8V_h]$$

$$0.488 V_h = 2.20 + 0.288 V_l \quad \Rightarrow \quad V_h \approx 30,93$$

$$0.32 V_l = 1.9 + 0.04 V_h \quad \Rightarrow \quad V_l \approx 6,904$$



Esempio di calcolo della funzione valore - IV



Devo specificare una policy (stocastica), robot più attivo:

$$s = \text{high} \quad [\text{Pr}(W) = 0.2 \quad \text{Pr}(S) = 0.8]$$

$$s = \text{low} \quad [\text{Pr}(W) = 0.2 \quad \text{Pr}(S) = 0.7 \quad \text{Pr}(R) = 0.1]$$

$$\begin{aligned} V_h &= 0.2 \times 1 \times [1 + 0.8V_h] + 0.8 \times 0.4 \times [3 + 0.8V_h] + 0.8 \times 0.6 \times [3 + 0.8V_l] \\ V_l &= 0.2 \times 1 \times [1 + 0.8V_l] + 0.7 \times 0.1 \times [3 + 0.8V_l] + 0.7 \times 0.9 \times [-3 + 0.8V_h] + \\ &\quad 0.1 \times 1 \times [0 + 0.8V_h] \end{aligned}$$

$$0.488 V_h = 2.20 + 0.288 V_l \quad \Rightarrow \quad V_h \approx -14.2$$

$$0.64 V_l = -0.8 + 0.44 V_h \quad \Rightarrow \quad V_l \approx -28.465$$

E' una policy peggiore.



Esempio di calcolo della funzione valore: ambiente generoso e stessa policy aggressiva



Stessa policy (stocastica) ma altro ambiente (robot che si scarica più difficilmente):

$$s = \text{high} \quad [\text{Pr}(W) = 0.2 \quad \text{Pr}(S) = 0.8]$$

$$s = \text{low} \quad [\text{Pr}(W) = 0.2 \quad \text{Pr}(S) = 0.7 \quad \text{Pr}(R) = 0.1]$$

$$\alpha = 0.4, \beta = 0.9, \gamma = 0.8, R_{\text{search}} = 6, R_{\text{wait}} = 1$$

$$\begin{aligned} V_h &= 0.2 \times 1 \times [1 + 0.8V_h] + 0.8 \times 0.4 \times [6 + 0.8V_h] + 0.6 \times 0.6 \times [6 + 0.8V_l] \\ V_l &= 0.2 \times 1 \times [1 + 0.8V_l] + 0.7 \times 0.9 \times [6 + 0.8V_l] + 0.7 \times 0.1 \times [-3 + 0.8V_h] + \\ &\quad 0.1 \times 1 \times [0 + 0.8V_h] \end{aligned}$$

$$0.584 V_h = 5.00 + 0.384 V_l \quad \Rightarrow \quad V_h \approx 13,12$$

$$0.784 V_l = -1.48 + 0.512 V_h \quad \Rightarrow \quad V_l \approx 6,942$$

Come utilizziamo V per determinare la policy ottimale?



Sommario



Determinazione della value function. Esempio.

Le equazioni di Bellman



Il modello markoviano



Il comportamento dell'ambiente è definito dallo stato: $S = \{s_j\}$
Per ogni stato l'agente sceglie un'azione: $a = a(s)$ $A = \{a_k\}$
Policy di un agente: $\pi(s, a)$ è quanto dobbiamo definire.

L'ambiente ha una evoluzione stocastica rappresentata da un MDP:

$$P_{s_t=s \rightarrow s_{t+1}=s' | a_t=a} = \Pr\{s_{t+1} = s' | s_t = s, a_t = a\}$$

Inoltre, ad ogni istante fornisce un reward immediato associato alla transizione, stimato all'istante t come:

$$R_{s_t=s \rightarrow s_{t+1}=s' | a_t=a} = E\{r_{t+1} = r' | s_t = s, a_t = a, s_{t+1} = s'\}$$

$$\forall s \in S; \forall a \in A$$



Probabilità composte



Probabilità degli eventi. It is a way of expressing knowledge or belief that an event will occur or has occurred. It ranges between 0 and 1.

Probabilità composta nel caso di eventi indipendenti:

Probabilità che due dadi diano 6 + 6 (probabilità **congiunta**)?

_ Probabilità che in una cassetta una mela sia gialla e bacata?

Probabilità che uno di due dati dia 2? $(p(1=2) * p(2 = 2) - p(1=2)p(2=2))$



Probabilità condizionata



Probabilità condizionata - consideriamo un mazzo di 40 carte:

vogliamo valutare quale sia la probabilità che una carta estratta a caso sia un re.

vogliamo valutare quale sia la probabilità che una carta estratta a caso sia un re, sapendo di avere estratto una figura.

$P(A)$ = probabilità che sia un re

$P(B)$ = probabilità che sia una figura

$$p(A | B) = 1/3$$

$$p(A | B) = p(A \cap B) / p(B) = 4/40 / 12/40 = 1/3$$



Probabilità Marginale



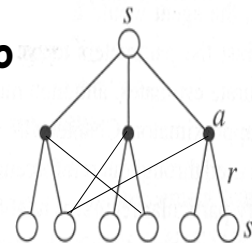
Probabilità totale di un certo evento.

$$p(A) = \sum_k p(A | B_k) \quad \text{Probabilità marginale.}$$



Reward stocastico

$$R_{s \rightarrow s' | a} = E\{r_{t+1} = r' | s_t = s, a_t = a, s_{t+1} = s'\}$$



reward = 3 – stocastico (da una distribuzione statistica)

È in realtà un valore condizionato in s e vale:

$$\Pr(\text{reward} = r | s') = \Pr(\text{reward} = r | s) / (\Pr(a)\Pr(s'))$$

Da Bayes:

$$\Pr(\text{reward} = r | s) = \Pr(\text{reward} = r | s') * \Pr(a | s) * \Pr(s' | s, a)$$

Questa è la probabilità congiunta di stato prossimo, azione e reward.



La value function



Nulla è detto sulla policy: dato uno stato, quale azione scegliere? In quale stato mi sposto?

Vogliamo costruire agenti lungimiranti.

State-Value function:

$$V^\pi(s) = E_\pi \{R_t \mid s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

Massimizzo la ricompensa a lungo termine, $V(\cdot)$. Dipende dalla policy:

A.A. 2010-2011

25/35

<http://homes.dsi.unimi.it/~borghese/>



Value function e modelli markoviani

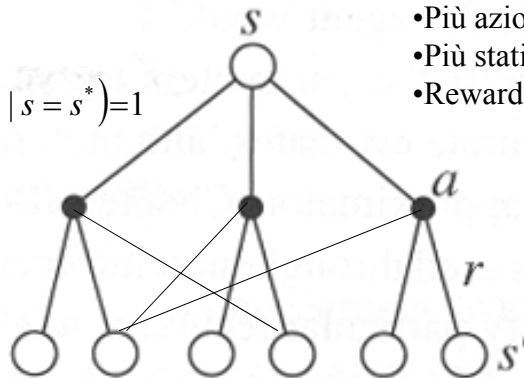


Anche la policy può essere stocastica.

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.

$$\sum_{j=1}^{N \text{ azioni}} \Pr(a_j \mid s = s^*) = 1$$



$$\sum_{k=1}^{N \text{ stati}} \Pr(s_{t+1} = s_k \mid s_t = s'; a_t = a_j) = 1$$

$$V^\pi(s) = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

A.A. 2010-2011

26/35

<http://homes.dsi.unimi.it/~borghese/>



Calcolo ricorsivo della Value function



$$V^\pi(s) = E_\pi \{R_t \mid s_t = s\} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

$$V^\pi(s') = E_\pi \{R_{t+1} \mid s_{t+1} = s'\} \quad \text{Relazione?}$$

$$V^\pi(s) = E_\pi \left\{ r_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k+1} \mid s_t = s \right\} =$$

$$V^\pi(s) = E_\pi \left\{ r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\}$$

Io termine

Illo termine

A.A. 2010-2011

27/35

<http://homes.dsi.unimi.it/~borghese/>



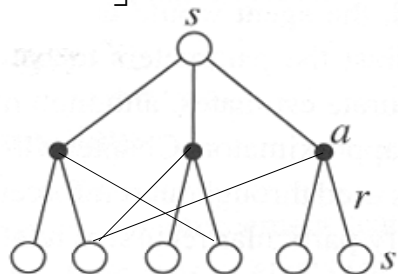
$V^\pi(s)$: primo termine



$$E_\pi \{r_{t+1} \mid s_t = s\} = \left[\sum_{a_j} \pi(a_j, s) \right] \sum_{s'} P_{s \rightarrow s' | a_j} [R_{s \rightarrow s' | a_j}]$$

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici nella transizione ad un passo



Visione Statistica: Probabilità di ottenere il reward: $R_{s \rightarrow s' | a_j}$ condizionata all'arrivare nello stato s' , che a sua volta è condizionata allo scegliere l'azione a_j .

A.A. 2010-2011

28/35

<http://homes.dsi.unimi.it/~borghese/>



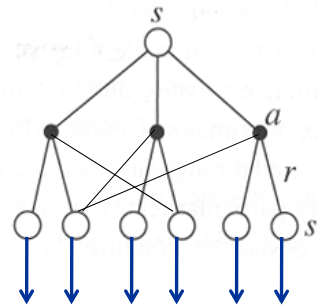
$V^\pi(s)$: secondo termine



$$V^\pi(s) = \dots + E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\} \quad V^\pi(s') = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s' \right\}$$

Per ogni stato devo valutare:

- Più azioni.
- Più stati prossimi
- Reward stocastici.



Nella valutazione dello stato s_t , sono considerati con un peso opportuno e scontati di γ , i reward a lungo termine che collezionerò a partire dai diversi stati s' .

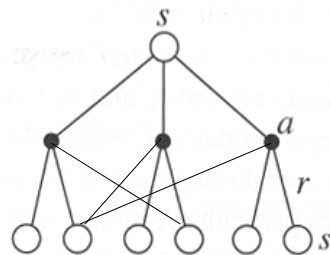


$V^\pi(s)$: secondo termine



$$V^\pi(s) = \dots + E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\} \quad V^\pi(s') = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s' \right\}$$

In s confluiranno i reward a lungo termine di tutti gli stati prossimi, s' , ciascuno pesato con la probabilità di passare da s a s' , ovvero sia, in termini statistici, condizionati alla realizzazione della transizione di stato, $s \rightarrow s'$.



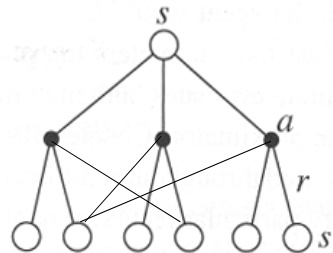
$$E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\} = \gamma \sum_{s'} E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s' \right\} \Pr(s_{t+1} = s' \mid s_t = s)$$



$V^\pi(s)$: secondo termine



$$E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\} = \gamma \sum_t E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_{t+1} = s'_l \right\} \Pr(s_{t+1} = s_k \mid s_t = s)$$



$$E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\} = \left[\sum_{a_j} \pi(a_j, s) \right] \sum_{s'} P_{s \rightarrow s' | a_j} [\gamma V^\pi(s')]]$$

A.A. 2010-2011

31/35

<http://homes.dsi.unimi.it/~borghese/>



Calcolo ricorsivo della Value function



$$V^\pi(s) = E_\pi \{ R_t \mid s_t = s \} = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\}$$

$$V^\pi(s') = E_\pi \{ R_{t+1} \mid s_{t+1} = s' \}$$

Legame?

Policy Next-state

$$P_{s \rightarrow s' | a} = \Pr \{ s_{t+1} = s' \mid s_t = s, a_t = a \}$$

$$V^\pi(s) = \sum_{a_j} \pi(a_j, s) \sum_{s'} P_{s \rightarrow s' | a_j} R_{s \rightarrow s' | a_j} + E_\pi \left\{ \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} \mid s_t = s \right\}$$

$$V^\pi(s) = \left\{ \sum_{a_j} \pi(a_j, s) \sum_{s'} \left\{ P_{s \rightarrow s' | a_j} \left[R_{s \rightarrow s' | a_j} + \gamma V^\pi(s'_l) \right] \right\} \right\}$$

Bellman's equation

A.A. 2010-2011

32/35

<http://homes.dsi.unimi.it/~borghese/>

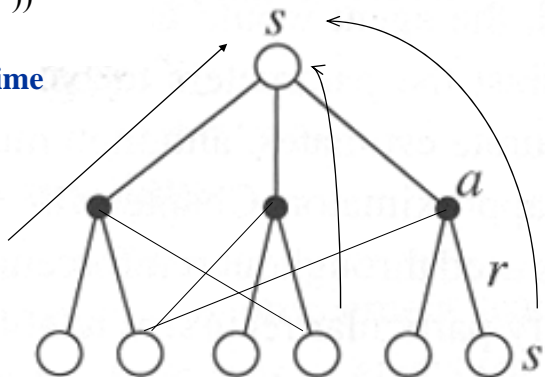


Osservazioni



$$V^\pi(s) = \text{funz}(V^\pi(s'))$$

Backwards in time



$$V^\pi(s) = \left\{ \sum_{a_j} \pi(a_j, s) \sum_{s_l'} \left\{ P_{s \rightarrow s_l' | a_j} \left[R_{s \rightarrow s_l' | a_j} + \gamma V^\pi(s_l') \right] \right\} \right\}$$

A.A. 2010-2011

33/35

<http://homes.dsi.unimi.it/~borghese/>



Confronto con il setting non associativo



	Setting non associativo	Setting associativo
Task	Azioni	Comportamenti (catena di azioni)
Reward	Reward istantaneo	Somma (scontata) dei reward collezionati lungo il task.
Max	Reward atteso sulla singola azione	Reward del comportamento
Orizzonte temporale del task	Finito (1 azione)	Finito / infinito per il singolo task
Policy	Stocastica	Stocastica
Stato	Non definito	Markoviano

A.A. 2010-2011

34/35

<http://homes.dsi.unimi.it/~borghese/>



Sommario



Determinazione della value function. Esempio.

Le equazioni di Bellman