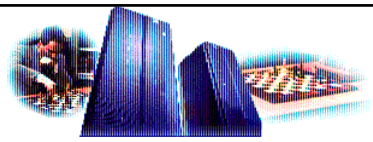
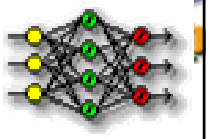




L'intelligenza biologica

Apprendimento con Rinforzo

Alberto Borghese
Università degli Studi di Milano
Laboratorio di Applied Intelligent Systems (AIS-Lab)
Dipartimento di Scienze dell'Informazione
borghese@dsi.unimi.it

Sommario

Il Reinforcement Learning

Il Reinforcement Learning con la Critica.

Il modello ASE / ACE.

I vari tipi di apprendimento

Supervisionato (learning with a teacher). Viene specificato per ogni pattern di input, il pattern desiderato in input.

Non-supervisionato (learning without a teacher). I neuroni verranno associati a pattern di ingresso contigui. Clustering. Mappe neurali.

Apprendimento con rinforzo (reinforcement learning, learning with a distal teacher). L'ambiente fornisce un'informazione del tipo success or fail.

Reinforcement learning

Nell'apprendimento supervisionato, esiste un "teacher" che dice al sistema quale è l'uscita corretta (learning with a teacher). Non sempre e' possibile.

Spesso si ha a disposizione solamente un'informazione giusto/sbagliato successo/fallimento.

Questa è un'informazione qualitativa ➔ ***learning with a critic.***

L'informazione disponibile si chiama ***segnale di rinforzo.*** *Non dà alcuna informazione su come aggiornare i pesi. Non è possibile definire una funzione costo o un gradiente.*

Obbiettivo: creare degli agenti "intelligenti" che abbiano una "machinery" per apprendere dalla loro esperienza.

Formalizzazione

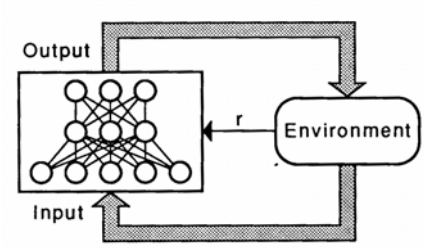
- Eseguire delle azioni sull'"ambiente" (Output)
- Osservare lo stato dell'"ambiente" (Inut).

Riceve un'informazione **puntuale** e **qualitativa** sul successo / (fallimento), della nostra azione, r.

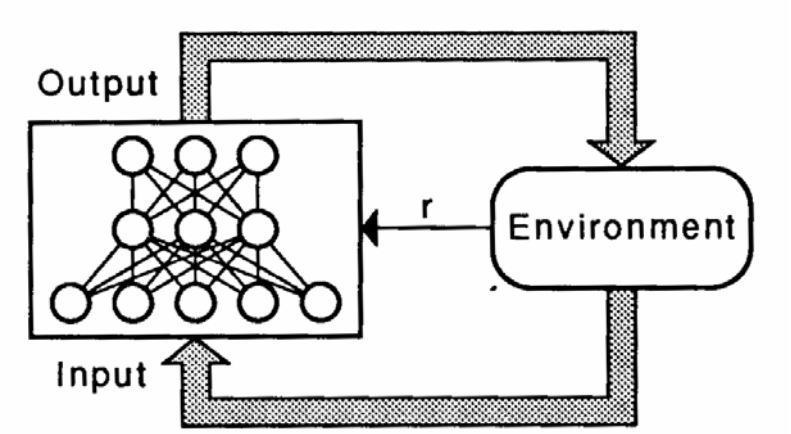
Imparare una politica di controllo (Output = f(Input)).

Come?

NB L'output interagisce con l'ambiente per generare un nuovo stato dell'"ambiente".



Reinforcement learning



Controllore o agente (e.g. NN): Funzione non-lineare multi-input / multi-output.
Ambiente: scalare, r (reward / penalty or success / fail).

I tue tipi di rinforzo

L'agente deve scoprire quale azione (policy) fornisca la ricompensa massima provando le varie azioni (trial-and-error).

"Learning is an adaptive change of behavior and that is indeed the reason of its existence in animals and man (K. Lorentz, 1977).

Rinforzo puntuale istante per istante, azione per azione (condizionamento classico).

Rinforzo puntuale "una-tantum" (condizionamento operante).

Il Condizionamento classico

La rete deve imparare una (o più) trasformazione tra input e output. Queste trasformazioni forniscono un comportamento che l'ambiente premia.

Il segnale di rinforzo è sempre lo stesso per ogni coppia input – output.

Esempio: risposte riflesse Pavolviane. Campanello (stimolo condizionante) prelude al cibo. Questo induce una risposta (salivazione). La risposta riflessa ad uno stimolo viene evocata da uno stimolo condizionante.

Stimolo-Risposta. Lo stimolo condizionante (campanello = input) induce la salivazione (uscita) in risposta al campanello.

Cf. Apprendimento Hebbiano.

Apprendimento con rinforzo di pattern di input/output – perceptrone con unità di attivazione lineari

$$J = E(\mathbf{w}) = \frac{1}{2}\sum_p \left[\sum_i \left(y_{ip}^D - y_{ip}\right)^2 = \frac{1}{2}\sum_i \left(y_{ip}^D - (\sum_j w_{ij} u_{jp}) \right)^2 \right]$$

$$\Delta w_{ij} = +\eta\left(y_i^D - y_i\right)u_j$$

δ rule (Hoff, 1960)

Possiamo supporre che le condizioni: $y_{ip} > y_{ip}^D$ e $y_{ip} < y_{ip}^D$ attivino l'apprendimento.

↓

$$\Delta w_{ij} = \eta\Theta\left(\left| y_i^D - y_i \right|\right)u_j$$

y : salivazione
u : campanello

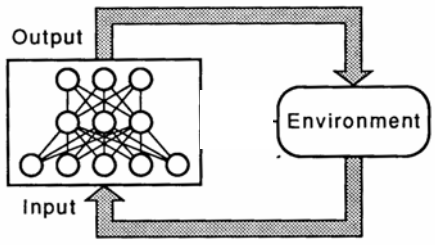
⇓

Il campanello induce la salivazione

Condizionamento operante

Reinforcement learning (operante).

Interessa un comportamento. Una sequenza di input / output che può essere modificata agendo sui parametri che definiscono il comportamento dell'agente. Il condizionamento arriva in un certo istante di tempo (spesso una-tantum) e deve valutare tutta la sequenza temporale di azioni, anche quelle precedenti nel tempo.



Proprietà del RL

L'ambiente o l'interazione può essere complessa.

Il rinforzo può avvenire solo dopo una più o meno lunga sequenza di azioni (delayed reward).

E.g. agente = giocatore di scacchi.
ambiente = avversario.

Problemi collegati:
temporal credit assignement.
structural credit assignement.

L'apprendimento non è più da esempi, ma dall'osservazione del proprio comportamento nell'ambiente.

Exploration vs Exploitation

Esplorazione (exploration) dello spazio delle azioni per scoprire le azioni migliori.

Le azioni migliori vengono scelte ripetutamente (exploitation) perchè garantiscono ricompensa (reward).

Occorre non interrompere l'esplorazione.

Occorre un approccio statistico per valutare le bontà delle azioni.

Exploration ed exploitation vanno bilanciate. Come?

RL operates on the whole task

RL considera il task complessivamente (non lo scompone in sotto-task o non modularizza il comportamento dell'agente).

Gli agenti hanno un goal esplicito che si realizza nell'interazione con l'ambiente.

L'ambiente può essere modellato in modo deterministico (lo scheletro + il supporto nel problema del controllo postura eretta) oppure con un certo grado di stocasticità (l'avversario nel gioco degli scacchi).

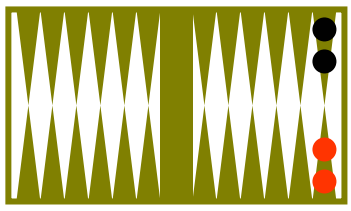
Back-gammon through RL (G. Tesauro, 1995)

Numero di situazioni:

- Configurazioni della scacchiera (10^{20})

Azioni:

- Mosse



Reward:

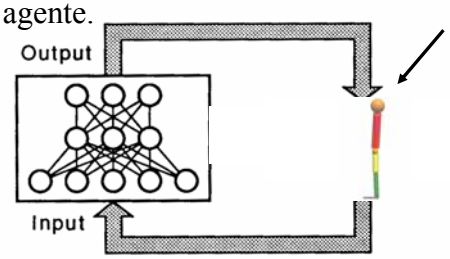
- +100 se vince
- - 100 se perde
- 0 per tutti gli altri stati

- Rete neurale allenata giocando 1,5 milioni di partite autonomamente.

Attualmente la macchina gioca a livello dei giocatori migliori.

Apprendimento del controllo della postura di un robot umanoide.

T(t) – le coppie articolare costituiscono l'output del nostro agente.



Da **ä(t)** tramite integrazione ottengo: **ȧ**(t) e **a**(t)

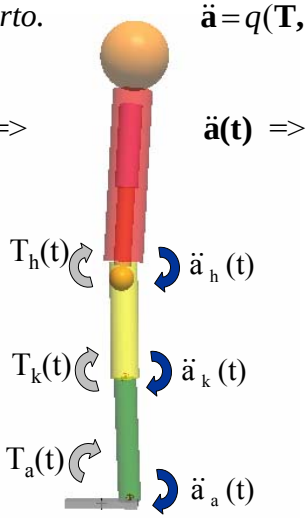
Considero lo stato, $\mathbf{s}(t)=[\dot{\mathbf{a}}(t);\mathbf{a}(t)]$ costituito da posizione e velocità dei segmenti. Lo stato coincide con l'input del nostro agente (e.g. NN).

"Environment" Sistema Dinamico. Rappresenta lo scheletro ed il supporto. $\ddot{\mathbf{a}}=q(\mathbf{T},\mathbf{a})$

T(t) =>

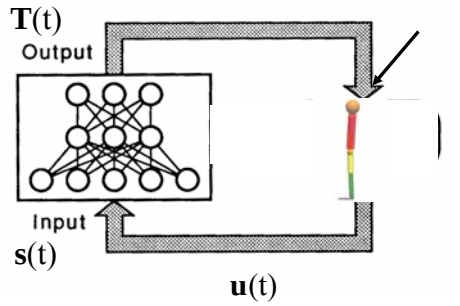
ä(t) =>

s(t)

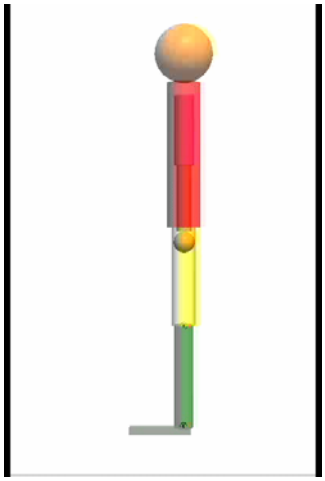


Comportamento iniziale (II)

"Environment"



ä(t) ➔ $\mathbf{s}(t)=[\dot{\mathbf{a}}(t);\mathbf{a}(t)]$



Esempi

Un giocatore di scacchi. Per ogni mossa ha informazione sulle configurazioni di pezzi che può creare e sulle possibili contro-mosse dell'avversario.

Una gazzella in 6 ore impara ad alzarsi e correre.

Un robot mobile deve decidere se collezionare altra spazzatura o ritornare a caricarsi. Deve decidere ciò in base alla sua esperienza di quanto tempo gli serva per tornare a caricarsi e quanto importante sia la spazzatura da collezionare.

Caratteristiche degli esempi

Interazione con l'ambiente. Un'azione porta l'agente in un nuovo stato di cui non si conosce compiutamente il valore.

Delayed reward. Per associare un valore ad uno stato occorre avere una stima di questo valore guardando una sequenza di azioni future ed il risultato di queste azioni con l'ambiente .

Miglioramento con l'esperienza. Miglioramento della valutazione delle situazione. Miglioramento delle scelte (di azioni). Conoscenza (implicita).

Problemi con orizzonte temporale finito o infinito.

Gli attori del RL

Policy. Descrive l'azione scelta dall'agente: mapping tra input (stato ambiente) e azioni. Funzione di controllo. Le policy possono avere una componente stocastica. Può essere implementata con una NN.

Reward function. Ricompensa immediata. Associata all'azione intrapresa in un certo stato. Può essere data al raggiungimento di un goal.

Value function. "Cost-to-go". Ricompensa a lungo termine. Somma dei reward + costi associati alle azioni scelte istante per istante.

Quale delle due è più difficile da ottenere?

Model of the environment. E' uno sviluppo relativamente recente. Da valutazione implicita dello svolgersi delle azioni future (trial-and-error) a valutazione esplicita mediante modello dell'ambiente della sequenza di azioni e stati futuri (planning).

Riassunto

•Reinforcement learning. I pesi vengono modificati, rinforzando le azioni che sono risultate buone a lungo termine.

•Self-discovery of a successful strategy (it does not need to be optimal!). La strategia (di movimento, di gioco) non è data a-priori ma viene appresa attraverso **trial-and-error**.

•Credit assignement (temporal and structural).

•Come possiamo procedere in modo efficiente nello scoprire una strategia di successo? Esplorazione dello spazio dei pesi?

Sommario

Il Reinforcement Learning

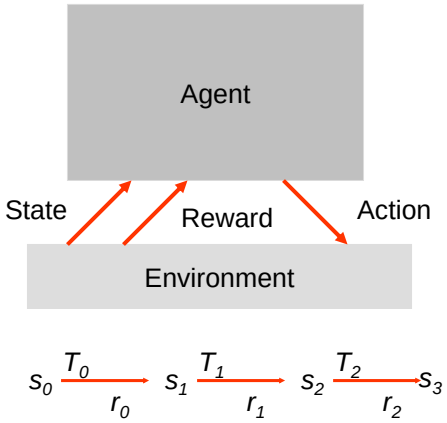
Il Reinforcement Learning con la Critica.

Il modello ASE / ACE.

La Funzione Rinforzo (Reward)

Viene ripetuto il ciclo:

- Eseguire delle azioni sul mondo {T}
- Osservare lo stato del mondo {s}.
- Osservare la ricompensa {r}.

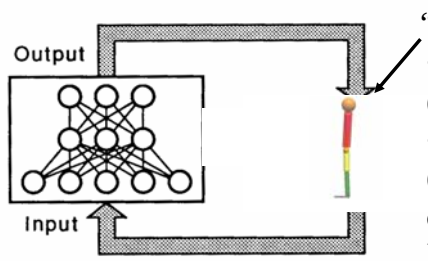


Imparare una politica di controllo (T = f(s)) tale che viene massimizzata la ricompensa totale ("life reward")

Da dove vengono gli $\{r_i\}$?

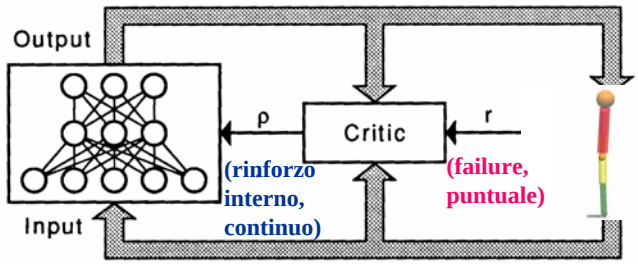
NB: Unsupervised learning. Delayed reward.

Reinforcement Learning (with a critic)



"Environment"

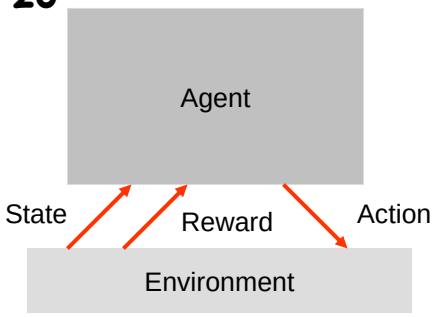
- r is the primary reinforcement (failure), scalare (reward).
- ρ is the secondary reinforcement (derivato dalla value function = cost-to-go, costo o ricompensa a lungo termine), scalare fornito con continuità nel tempo.



Lo schema dell'apprendimento con rinforzo

Viene ripetuto il ciclo:

- Eseguire delle azioni sul mondo {T}
- Osservare lo stato del mondo {s}.
- Osservare la ricompensa {r}.



Imparare una politica di controllo (T = f(s)) tale che viene massimizzata la ricompensa totale ("life reward")

$$s_0 \xrightarrow[r_0]{T_0} s_1 \xrightarrow[r_1]{T_1} s_2 \xrightarrow[r_2]{T_2} s_3$$

Imparare una valutazione degli stati in funzione al loro "grado di rischio" o "grado di ricompensa" che promettono.

Come posso valutare la ricompensa a lungo termine?

- Ho bisogno di una funzione (Value function) che per ogni stato presente, in funzione della sequenza di azioni (policy) che prevedo di scegliere in futuro, mi possa dire quanto mi costa, o quanto è vantaggiosa la policy di controllo utilizzata nell'istante presente *(T(t) = f(s(t)))*.

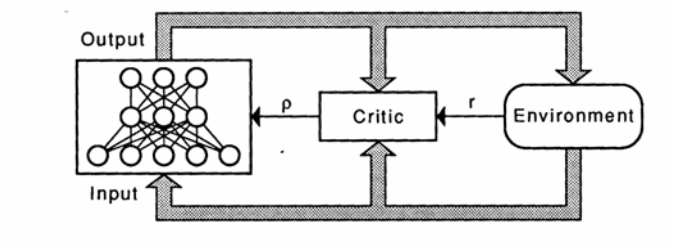
- La Value function è una funzione che rappresenta la mappa di rischio.

Struttura della critica

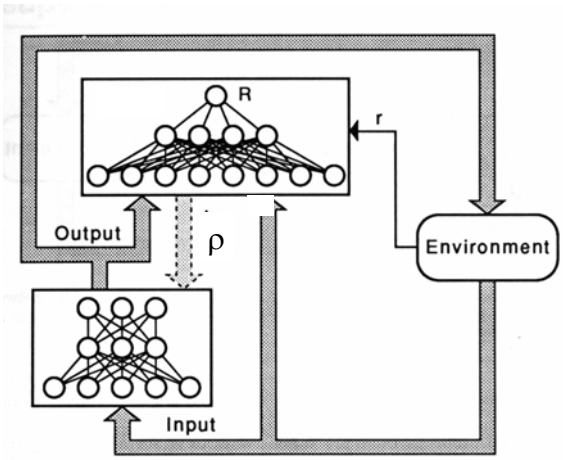
Per ogni istante t, la value function, o mappa di rischio, J(t) = J(**s**(t)), è una funzione dello **stato**.

J(.) può essere rappresentato da una funzione non-lineare, derivabile (e.g. NN).

La critica impara una value function per ogni stato, ed invia al controllore un segnale di rinforzo interno, o secondario: ρ(t).



Da dove nasce la value function?



•Deve essere appresa anch'essa.

•Deve trasformare lo scalare r puntuale (una-tantum), in un secondo salare ρ, fornito con continuità nel tempo.

•**Seconda rete neurale specializzata nell'apprendimento della value function o mappa di rischio.**

Sommario

Il Reinforcement Learning.

Il Reinforcement Learning con la Critica.

Il modello ASE / ACE.

Un'implementazione di RL (ACE/ASE)

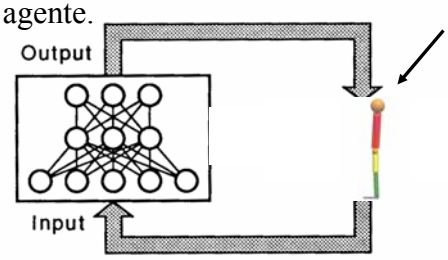
A. Barto, R. Sutton and C:W. Anderson, Neuron-like Adaptive Elements That Can Solve Difficult Learning Control Problems, IEEE Trans. Systems, Man and Cybernetics, 1983.

ASE – Adaptive Search Element – Controllore.

ACE – Adaptive Critic Element – Critica.

Apprendimento del controllo della postura di un robot umanoide.

T(t) – le coppie articolare costituiscono l'output del nostro agente.



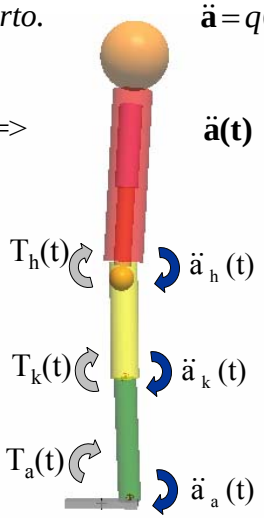
Da $\ddot{\mathbf{a}}(t)$ tramite integrazione ottengo: $\dot{\mathbf{a}}(t)$ e $\mathbf{a}(t)$

Considero lo stato, $\mathbf{s}(t)=[\dot{\mathbf{a}}(t);\mathbf{a}(t)]$ costituito da posizione e velocità dei segmenti. Lo stato coincide con l'input del nostro agente (e.g. NN).

"Environment" Sistema Dinamico. Rappresenta lo scheletro ed il supporto. $\ddot{\mathbf{a}}=q(\mathbf{T},\mathbf{a})$

$\mathbf{T}(t)$ => $\ddot{\mathbf{a}}(t)$ =>

$\mathbf{s}(t)$



Rappresentazione a box delle variabili di stato

Le variabili sono codificate a **box** (intervalli disgiunti).

Orientamento del polpaccio rispetto ad un asse verticale $\vartheta : 0,\pm4,\pm12,\pm24$ deg
Velocità angolare del polpaccio $\dot{\vartheta} : \pm50,\pm\infty$ deg/s

Orientamento della coscia rispetto ad un asse verticale $\omega : 0,\pm4,\pm12,\pm24$ deg
Velocità angolare della coscia $\dot{\omega} : \pm50,\pm\infty$ deg/s

Orientamento del tronco rispetto ad un asse verticale $\varphi : 0,\pm4,\pm12,\pm24$ deg
Velocità angolare del tronco $\dot{\varphi} : \pm50,\pm\infty$ deg/s

Altra possibilità: fuzzy set. CMAC.

Modellazione del controllore con RL

Suppongo s(t) = 0 se il sistema non si trova in quel particolare stato, oppure s(t) = 1 viceversa.

Il segnale di rinforzo esterno (reward) r = -1 nel momento della failure, altrimenti r = 0.

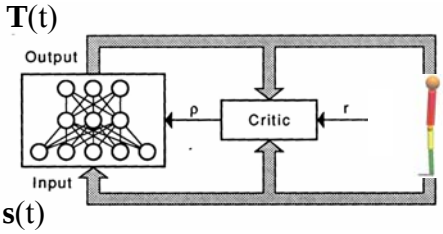
Considero che la critica mi fornisca uno scalare graduato che rappresenta il mio rinforzo interno. E' estratto dalla stima del "long-time reward" = "cost-to-go" = Value function.

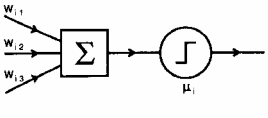
Considero che il controllore fornisca uno scalare -1 o 1 per ciascuna delle variabili di controllo (bang-bang controller) – *Policy*.

Environment deterministico (sistema muscolo-scheletrico + supporto).

Struttura del controllore

$$T_j(t) = \Theta\left(\sum_i w_{ij} s_i(t) + noise(t)\right)$$





Noise(t) – ha il ruolo di incoraggiare l'esplorazione dello spazio.

I pesi $\{w_{ij}\}$ variano per effetto dell'apprendimento.

L'eleggibilità

$$T_j(t) = \Theta\left(\sum_i w_{ij} s_i(t) + noise(t)\right)$$

$$e_{ij}^{\,c}(t+1) = \delta e_{ij}^{\,c}(t) + (1-\delta) T_j(t) s_i(t) \qquad \delta < 1$$

Se uno stato $s_i(t)$ non viene visitato ($s_i(t) = 0$), la eleggibilità dell'azione scelta per quello stato decresce esponenzialmente.

Se uno stato $s_i(t)$ viene visitato ($s_i(t) = 1$), si modifica il valore di tutti i pesi associati a quella unità di input:

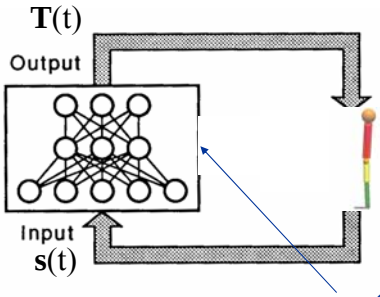
se $T_j(t)$ rimane dello stesso segno, la sua eleggibilità tende a $T_j * s_i$.
se $T_j(t)$ cambia spesso segno, la sua eleggibilità tende a 0.

La eleggibilità *aggiunge perciò la dimensione temporale al prodotto $T_j * s_i$: questo viene considerato valido solamente se si ripete nel tempo e se si ripete uguale (e.g. Torque positivo o negativo associato allo stato s_i).*

Aggiornamento del controllore

$$T_j(t) = \Theta\left(\sum_i w_{ij} s_i(t) + noise(t)\right)$$

$$e_{ij}^{\ c}(t+1) = \delta e_{ij}^{\ c}(t) + (1-\delta) T_j(t) s_i(t)$$



$$\Delta w_{ij}^{\ c} = \alpha \rho(t) e_{ij}(t)$$

e_{ij}(t) – eleggibilità del peso ij.

ρ(t) – reward (interno)

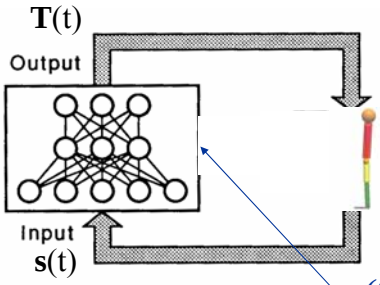
Il rinforzo, ρ(t), decide l'intensità dell'aggiornamento del peso ij al tempo t.

Osservazioni

L'aggiornamento qui dipende dalla storia del sistema.

e_{ij}(t) – eleggibilità del peso ij.

$$\Delta w_{ij}^{\ c} = \alpha \rho(t) e_{ij}(t)$$



Nel caso del perceptrone era:

$$\Delta w_{ij} = +\eta\left(y_i^D - y_i\right)u_j$$

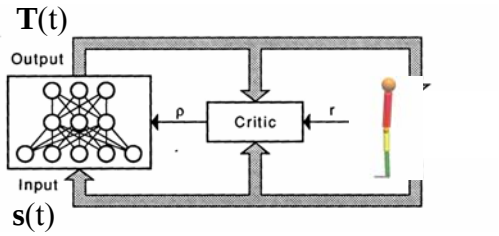
ρ(t) – reward (interno)

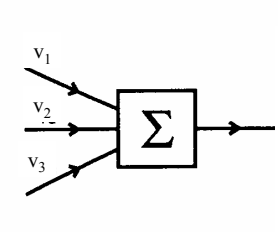
NB Lo structural credit assignement è risolto dall'eleggibilità.

Struttura della critica

$$p(t) = \left(\sum_i v_i s_i(t)\right)$$

p(t) – mappa di rischio, stima long-term reward = cost-to-go = value function..



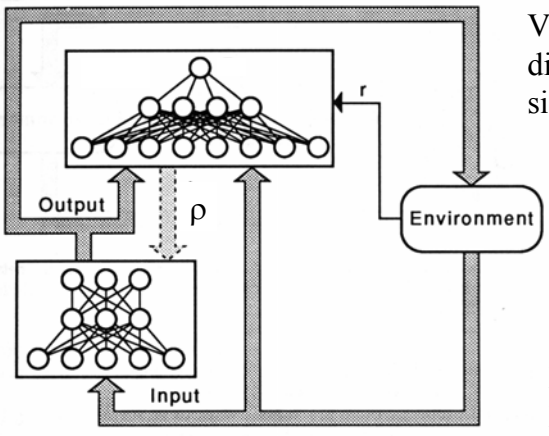


Razionale: Se passo da uno stato più rischioso ad uno stato meno rischioso, l'azione va rinforzata.

I pesi $\{v_i\}$ variano per effetto dell'apprendimento.

La value function, p(t), e la mappa di rischio

Due passi:



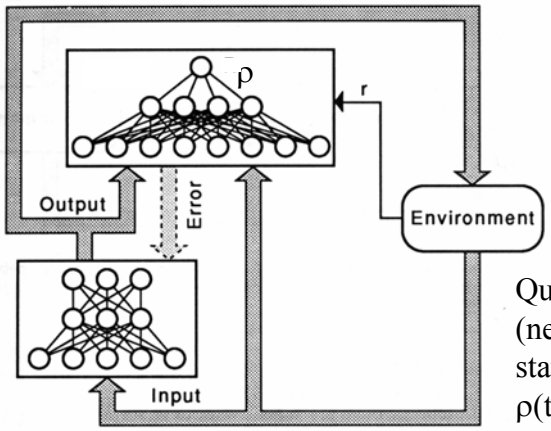
Viene calcolato per ogni istante di tempo, lo stato di rischio del sistema, p(t):

$$p(t) = \left(\sum_i v_i s_i(t)\right)$$

Dallo stato di rischio attuale e dallo stato di rischio precedente (e dal rinforzo puntuale, r), determino il rinforzo interno, ρ(t).

Generazione del rinforzo interno

$$\rho(t) = r(t) + \gamma p(t) - p(t-1) \quad 0 < \gamma \leq 1$$



Fino a quando il controllore riesce a mantenere la postura eretta (nessun fallimento, r = 0), ρ(t) è **positivo**, quando il sistema passa da uno stato a più alto grado di rischio ad uno con un grado di rischio inferiore.

Quando arriva il reinforcement (negativo), r = -1. Non ci sono stati associati, per cui p(T) = 0. ρ(t) diventa **negativo**:

$\rho(t) = -1 - p(t-1)$.

L'eleggibilità nell'apprendimento della mappa di rischio.

$$p(t) = \left(\sum_i v_i s_i(t)\right) \qquad \rho(t) = r(t) + \gamma p(t) - p(t-1) \quad 0 < \gamma \leq 1$$

Eligibility di uno stato $s_i(t)$ dipende da quante volte lo stato è stato visitato nel passato. Uno stato sempre visitato avrà eligibility massima:

$$e_i^r(t+1) = \lambda e_i^r(t) + (1-\lambda) s_i(t) \qquad \lambda < 1$$

λ regola la plasticità. Maggiore è λ , maggiore il tempo di decadimento, minore è λ , maggiore è la variazione di eleggibilità dovuta a visite recenti dello stato i.

Apprendimento della mappa di rischio, p(t)

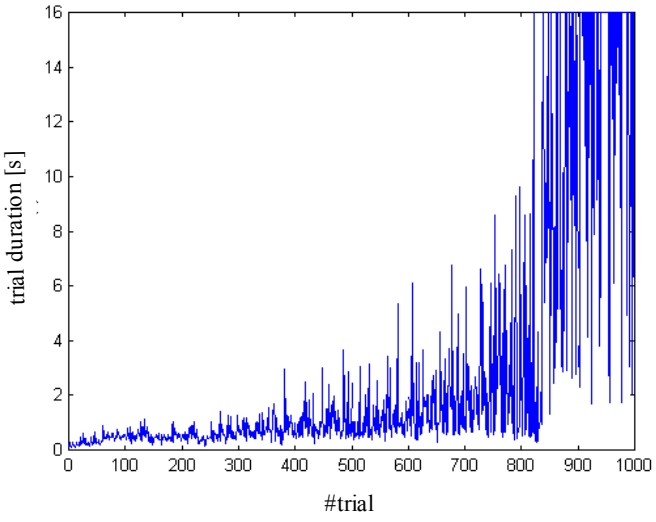
$$p(t) = \left(\sum_i v_i(t) s_i(t)\right) \qquad \rho(t) = r(t) + \gamma p(t) - p(t-1) \quad 0 < \gamma \leq 1$$

Aggiorno la mappa di rischio rinforzando quei pesi associati agli stati visitati più di recente, più eleggibili:

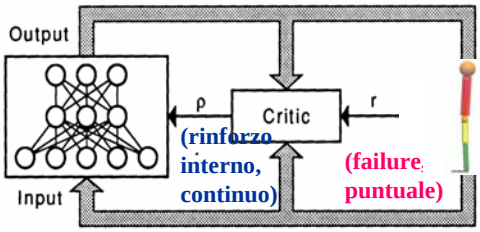
$$\Delta v_i = \beta \rho(t) e_i^r(t)$$

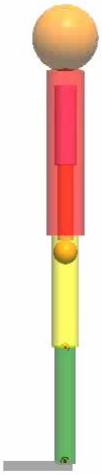
p(t) può avere segno positivo o negativo e regola l'incremento o decremento dei pesi che esprimono la value function associata allo stato.

Curva di apprendimento



Apprendimento





La Stanza Cinese (J. Searle, 1980)

La persona (CPU).
Un libro di regole (Il programma).
Un pacco di fogli (la memoria).

Il calcolatore potrebbe dimostrare di essere intelligente al test di Turing, senza comprendere nulla.
Il signore nella stanza cinese riceve in ingresso dei simboli che manipola secondo regole a lui ignote e poi fornisce le risposte.

Lui non conosce il cinese!

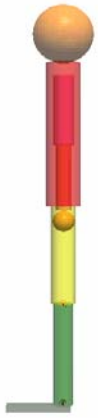
Riassunto sull'apprendimento con rinforzo

Necessita di una *critica*, che trasforma il segnale scalare di rinforzo (puntuale) in un segnale scalare temporale, r(T) -> ρ(t).

La critica analizza le coppie input/output ed impara una mappa di rischio.

Utilizza questa mappa di rischio per fornire un segnale di rinforzo interno al controllore.

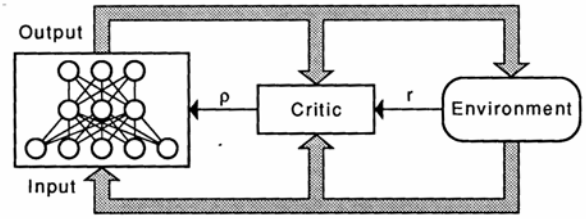
Il segnale di rinforzo viene utilizzato per aggiornare i pesi associati a buone scelte (buone azioni) con un meccanismo Hebbiano, dove la valutazione avviene lungo la dimensione temporale.



RL con la critica

La critica deve valutare il funzionamento del controllore in un modo che sia: **appropriato** per l'obbiettivo del controllo e sufficientemente **informativo** perché il controllore apprenda.

Determinare **come variare i pesi** del controllore in modo da migliorare le prestazioni, misurate dalla critica.



Sommario

Il Reinforcement Learning.

Il Reinforcement Learning con la Critica.

Il modello ASE / ACE.