Assignment 5 - Reinforcement Learning

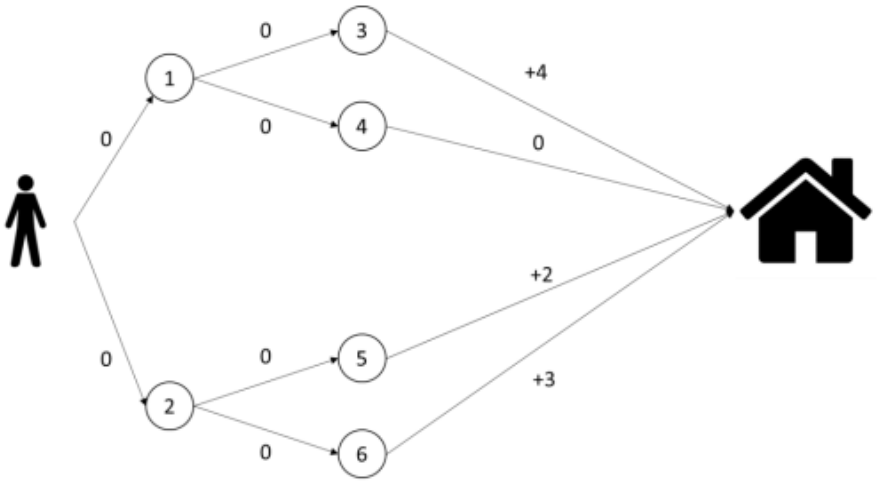
Dario Carolla mat. 807547

June 14, 2019

Considerando la seguente figura, sono presenti otto diversi stati:

- Uno stato iniziale 0 (dove si trova l'uomo)
- Sei stati intermedi (da 1 a 6)
- Uno stato finale F (dove si trova la casa)

In tutti gli stati, ad eccezione di quelli che portano allo stato finale, è necessario prendere una decisione: andare a destra o a sinistra.



È fornita, inoltre, la seguente policy iniziale π_0:

- $0 \rightarrow 2$
- $2 \rightarrow 5$
- $1 \rightarrow 4$
- Dagli stati 3, 4, 5 e 6 è possibile andare solo in una direzione

Lo scopo del progetto è risolvere il *Markov decision process (MDP)*. Un MDP è definito da:

- S - lo spazio finito dello stato
- A - lo spazio finito delle azioni
- T - matrice di transizione per processi deterministici, $T(s, a) \rightarrow s'$
- R(s,a) - Reward che, in questo caso, è data dal peso degli archi
- π una politica che associa uno stato a un'azione $\pi : S \rightarrow A$

Risolvere un MDP significa trovare una politica ottimale π^*, identificando così la migliore azione da eseguire in un dato stato, in modo tale da ottenere il massimo valore di una ricompensa. In questo caso, lo scopo è massimizzare la reward. Inizialmente, viene calcolata la funzione del valore di ogni stato considerando la politica attuale nel seguente modo:

$$V(s)^{\pi} = R(s, \pi(s)) + \gamma \sum_{s' \in S} T(s, \pi(s), s')V^{\pi}(s')$$

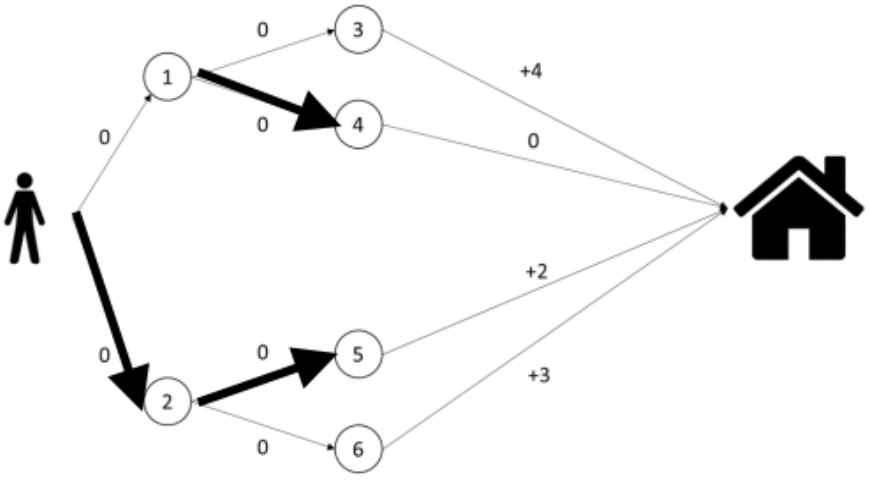
Il valore ottimale dello stato è dato da:

$$V^*(s) = max_{a \in A}[R(s, a) + \gamma \sum_{s' \in S} T(s, a, s')V^*(s')]$$

Infine, la politica ottimale viene calcolata tramite l'equazione di Bellman:

$$\pi^*(s) = argmax_{a \in A}[R(s, a) + \gamma \sum_{s' \in S} T(s, a, s')V^*(s')]$$

Nella seguente immagine, con i rami più marcati è stata rappresentata la politica attuale:

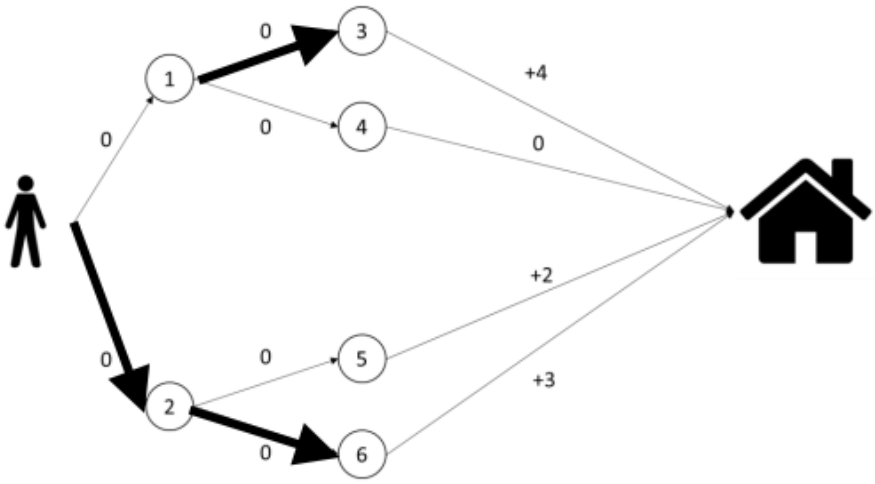


In questo caso la reward ottenuta, con il cammino $0 \to 2 \to 5 \to F$, è 2. Tenendo in considerazione la politica attuale vengono calcolati i valori ottimali degli stati. Per semplicità si considera $\gamma = 1$:

- $V(F) = 0$
- $V(3) = max\{4\}$ (solo un'azione possibile $3 \to F$)
- $V(4) = max\{0\}$ (solo un'azione possibile $4 \to F$)
- $V(5) = max\{2\}$ (solo un'azione possibile $5 \to F$)
- $V(6) = max\{3\}$ (solo un'azione possibile $6 \to F$)
- $V(1) = max\{0 + \gamma 0, 0 + \gamma 4\} = 4$ (azione $1 \to 3$)
- $V(2) = max\{0 + \gamma 2, 0 + \gamma 3\} = 3$ (azione $2 \to 6$)
- $V(0) = max\{0 + \gamma 0, 0 + \gamma 2\} = 2$ (azione $0 \to 2$)

La nuova politica risulta essere:

- $1 \to 3$
- $2 \to 6$
- $0 \to 2$



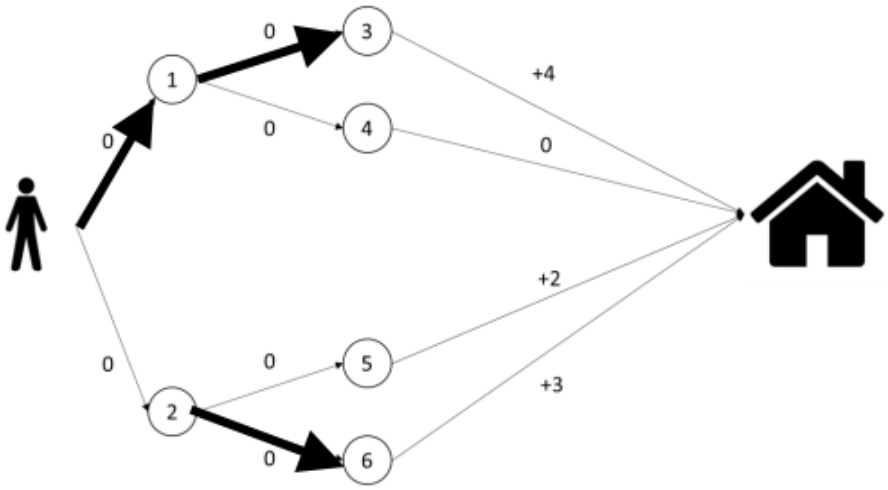
Con tale politica la reward, data dal cammino $0 \to 2 \to 5 \to F$, è pari a 3. Di seguito vengono calcolati i nuovi valori ottimali per ogni stato considerando la nuova politica:

- $V(F) = 0$

- $V(3) = max\{4\}$ (solo un'azione possibile $3 \rightarrow F$)
- $V(4) = max\{0\}$ (solo un'azione possibile $4 \rightarrow F$)
- $V(5) = max\{2\}$ (solo un'azione possibile $5 \rightarrow F$)
- $V(6) = max\{3\}$ (solo un'azione possibile $6 \rightarrow F$)
- $V(1) = max\{0 + \gamma 0, 0 + \gamma 4\} = 4$ (azione $1 \rightarrow 3$)
- $V(2) = max\{0 + \gamma 2, 0 + \gamma 3\} = 3$ (azione $2 \rightarrow 6$)
- $V(0) = max\{0 + \gamma 3, 0 + \gamma 4\} = 4$ (azione $0 \rightarrow 1$)

La nuova politica risulta:

- $1 \rightarrow 3$
- $2 \rightarrow 6$
- $0 \rightarrow 1$



La politica attuale risulta avere una $Reward = 4$ con il cammino $0 \rightarrow 1 \rightarrow 3 \rightarrow F$. In particolare, la politica attuale risulta essere quella ottimale, in quanto, per ogni nodo in cui è possibile effettuare una decisione, viene effettuata quella con la reward maggiore, in questo modo è stato raggiunto l'obbiettivo di massimizzare la reward.