

Invatarea prin recompensa

Value Iteration

Algoritmul Value Iteration

Calculeaza utilitatea starilor (recompensa pe termen lung):

$$U_{t+1}(S_i) = R(S_i) + \delta \max_a \sum_j T(S_i, a, S_j) * U_t(S_j)$$

Intrare: T(model al tranzitiilor), R (recompensa starilor)

Iesire: U (utilitatea fiecărei stări)

```
{
    U = utilitatea fiecărei stări (initial identica cu R)
    U' = utilitatea fiecărei stări (la pasul următor - initial identica cu R)
    repeta
    {
        U ← U'
        pentru fiecare stare Si
            U'(Si) = R(Si) + δ maxa ∑j T(Si, a, Sj) * U(Sj)
    }pana cand distanta(U, U') < ε
    return U
}
```

$$\text{distanța}(U, U') = \frac{1}{S} \sqrt{\sum_{i=1}^S (U(S_i) - U'(S_i))^2}, S = \text{numărul de stări}$$

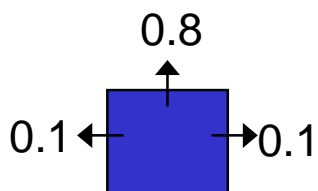
$$\text{Politica optimă } \pi^*(S_i) = \arg \max_a \sum_j (R(S_i) + T(S_i, a, S_j) * U(S_j))$$

Exemplu: grid 3 x 4

3				+1
2				-1
1				
	1	2	3	4

- Doua stari terminale cu recompensele +1 si -1
- Toate celelalte stari au recompensele -0.04
- Deplasarea se poate face in directiile Nord/Sud/Est/Vest

Exemplu: Modelul mediului T - deplasare in directia N



Exemplu de calcul a utilitatii pentru starea S ($\delta = 1$)

a)										
<table><tr><td></td><td>U=10</td><td></td></tr><tr><td>U=5</td><td>S (R=1)</td><td>U= -8</td></tr><tr><td></td><td>U=1</td><td></td></tr></table>		U=10		U=5	S (R=1)	U= -8		U=1		$U_{t+1}(S) = 1 + \max(0.8*5+0.1*10+0.1*1, (\leftarrow)$ $0.8*10+0.1*(-8)+0.1*5, (\uparrow)$ $0.8*(-8)+0.1*10+0.1*1, (\rightarrow)$ $0.8*1+0.1*5+0.1*(-8) (\downarrow))=$ $= 1 + \max (5, 1 (\leftarrow), 7.7 (\uparrow), -5.3 (\rightarrow), 0.5 (\downarrow))$ $= 1 + 7.7 = 8.7 (\uparrow)$
	U=10									
U=5	S (R=1)	U= -8								
	U=1									
b)										
<table><tr><td>U=5</td><td>U=10</td></tr><tr><td>S (R=1, U_t=1)</td><td>U=1</td></tr></table>	U=5	U=10	S (R=1, U _t =1)	U=1	$U_{t+1}(S) = 1 + \max(0.8*1+0.1*5+0.1*1, (\leftarrow)$ $0.8*5+0.1*1+0.1*1, (\uparrow)$ $0.8*1+0.1*1+0.1*5, (\rightarrow)$ $0.8*1+0.1*1+0.1*1 (\downarrow))=$ $= 1 + \max (1.4 (\leftarrow), 4.2 (\uparrow), 1.4 (\rightarrow), 1 (\downarrow))$ $= 1 + 4.2 = 5.2 (\uparrow)$					
U=5	U=10									
S (R=1, U _t =1)	U=1									