2018101036
MDL - Assgn 2 (Part 1)
BHAVESH SHUKLA

Using value iteration algorithm, we have :-

$U_t(i) = \max [R(i,a) + \gamma \sum_j P(j/ia) \times U_t(j)]$

where $U_t(i)$ is utility of state i on time t.

This eq^n is called Bellman Update Eq^n

States : | S_0 | S_1 | S_2 | S_3 +10 | where S_3 is absorbant state

Actions → [l, r] l = left & r = right

l → move left with p = 0.8, right with p = 0.2

r → move left with p = 0.2, left with p = 0.8

Bellman factor = 0.01 = δ

Gamma = γ = 0.25

t_0 → Iteration 0

t_0 = | 0 | 0 | 0 | +10 |

t_1 → Iteration 1

$S_0 = \max(-1+\gamma(0\times0.8+0\times0.2), -1+\gamma(0\times0.8+0\times0.2))$

$S_0 = \max(-1,-1) = -1$

$S_1 = \max(-1+\gamma(0\times0.8+0\times0.2), -1+\gamma(0\times0.8+0\times0.2))$

$S_1 = \max(-1,-1) = -1$

$S_2 = \max(-1+\gamma(0\times0.8+10\times0.2), -1+\gamma(10\times0.8+0\times0.2))$

$= \max(-0.5, 1) = 1$

$\Rightarrow t_1 =$ | -1 | -1 | +1 | +10 | , $\delta = 1$ — Deterministic

t_2 → Iteration 2

$S_0 = \max(-1+\gamma(-1\times0.8 + -1\times0.2), -1+\gamma(-1\times0.2 + -1\times0.8))$

$= \max(-1+\gamma(-1), -1+\gamma(-1))$

$= -1-\gamma = -1.25$

$S_1 = \max(-1+\gamma(+1\times0.8 + 1\times0.2), -1+\gamma(\cdots))$

$= \max(-1+\gamma(-0.6), -1+\gamma(0.6))$

$= -1+\gamma(0.6) = -0.85$

$S_2 = \max(-1+\gamma(-1\times0.8+10\times0.2), -1+\gamma(10\times0.8+-1\times0.2))$

$= \max(-1+\gamma(1.2), -1+\gamma(7.8))$

$= -1+\gamma(7.8) = +0.95$

$\Rightarrow t_2 =$ | -1.25 | -0.85 | +0.95 | +10 | , $\delta = 0.25$

t_3 → Iteration 3

$S_0 = \max(-1+\gamma(-1.25\times0.8 + -0.85\times0.2), -1+\gamma(-0.85\times0.8 + -1.25\times0.2))$

$= \max(-0.21, -1.2325)$ $\max(-1+\gamma(-1.61), -1+\gamma(-0.945))$

$= -1+\gamma(-0.945)$

$= -1.2325$

$S_1 = \max(-1+\gamma(-1.25\times0.8 + -0.95\times0.2), -1+\gamma(0.95\times0.8 + -1.25\times0.2))$

$= \max(-1.2975, -0.8725) = -0.8725$

$S_2 = \max(-1+0.25(-0.85\times0.8+10\times0.2), -1+\gamma(10\times0.8 + -0.85\times0.2))$

$S_2 = \max(-1+0.23, -1+0.25(8-0.17))$

$S_2 = \max\{-0.67, 0.9575\} = 0.9575$

$\Rightarrow t_3 = \boxed{-1.2325 \mid -0.8725 \mid 0.9575 \mid 10}$, $\delta = 0.0225$

$t_4 \rightarrow$ Iteration 4

$S_0 = \max\{-1 + 0.25(0.8 \times (-1.233) + 0.2 \times (-0.873)),$
$-1 + \gamma(0.2 \times (-1.233) + 0.8 \times (-0.873))\}$
$= \max(-1 + 0.25(-1.61, -0.945)) = -1 + 0.25(-0.945)$
$= -1.236$

$S_1 = \max\{-1 + 0.25(0.8 \times (-1.233) + 0.2(0.958)),$
$-1 + 0.25(0.2 \times (-1.233) + 0.8 \times (0.958))\}$
$= \max(-1 + 0.25 \times (-0.794), -1 + 0.25 \times 0.52)$
$= -0.873$

$S_2 = \max(-1 + \gamma(0.8 \times (-0.873) + 0.2 \times 10),$
$-1 + 8(0.2 \times (-0.873) + 0.8 \times 10))$
$= 0.958$

$\Rightarrow t_4 = \boxed{-1.236 \mid -0.873 \mid 0.958 \mid 10}$

$\delta = 0.003625$

Max function results in move right

Ans) $\Rightarrow$ Policy π = {Move Right, Move Right, Move Right, No Action}

1) Reward is always 10
2) Movement on extremes is same as staying in same all.