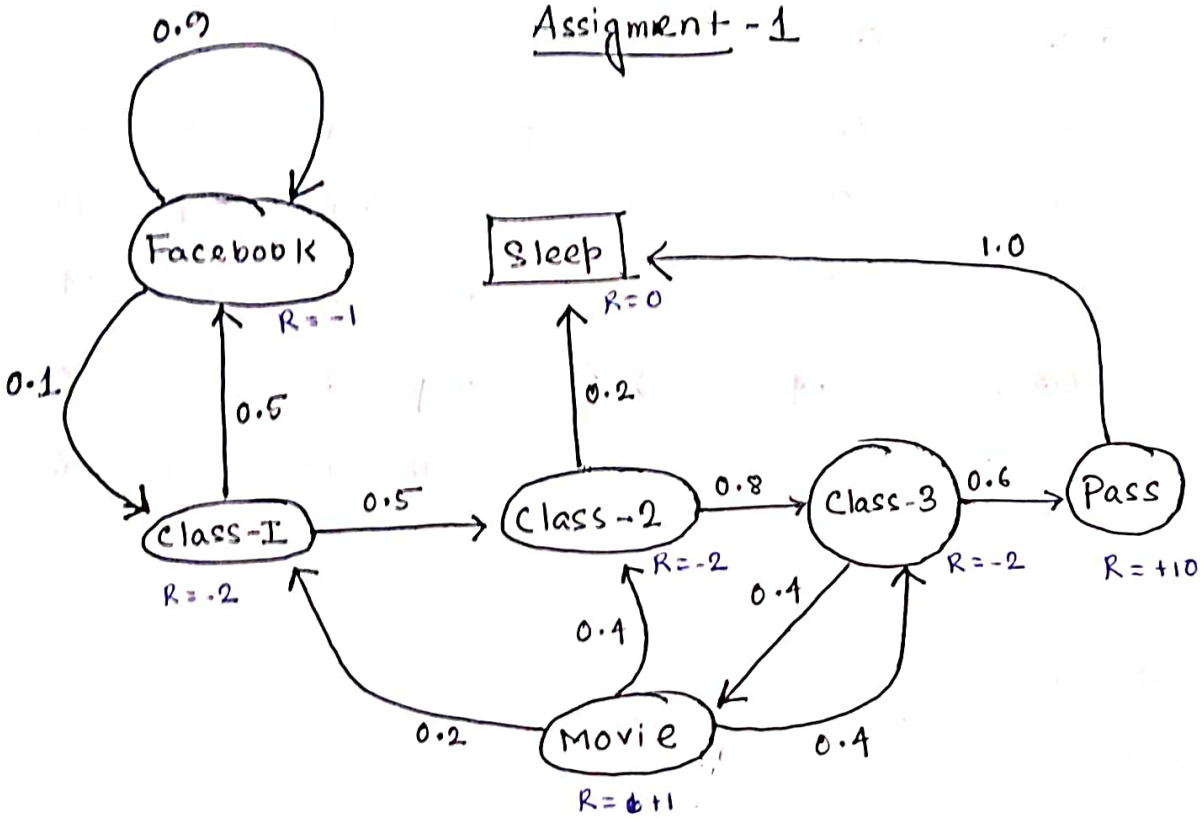
Name - Anirban Dey
Reg NO - B2230019

Subject - Reinforcement Learning.

## Assignment - 1




→ States :-
- Facebook
- Class 1
- Class 2
- Class 3
- Pass
- Movie
- Sleep

Action : The actions corresponding to transitions between states. In this case the actions are just moving from one state to another

Transition Probabilities :-.

| | Facebook | Class-1 | Class-2 | Class-3 | Pass | Movie | Sleep |
|---|---|---|---|---|---|---|---|
| • Facebook | 0.9 | 0.1 | 0 | 0 | 0 | 0 | 0 |
| Class-1 | 0.5 | 0 | 0.5 | 0 | 0 | 0 | 0 |
| Class-2 | 0 | 0 | 0 | 0.8 | 0 | 0 | 0.2 |
| Class-3 | 0 | 0 | 0 | 0 | 0.6 | 0.4 | 0 |
| Pass | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Movie | 0 | 0.2 | 0.4 | 0.4 | 0 | 0 | 0 |
| Sleep | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Rewards :-.

Facebook → -1

Class 1 → -2

Class 2 → -2

Class 3 → -2

Pass → +10

Movie → +1

Sleep → 0.

Step-1 We have to first initialize the value function $v(s)$ for all States. We can start with $v(s)=0 \quad \forall s \in S$.

v (Facebook) = 0 value (Movie) = 0

v (class-1) = 0 value (sleep) = 0

v (class-2) = 0

v (class-3) = 0

v (Pass) = 0

## eration - 1

$V_1$ (Facebook) $= (-1) + [0.9 \times 0 + 0.1 \times 0] = -1$

$V_1$ (Class 1) $= (-2) + (0.5 \times \overset{-1}{0} + 0.5 \times 0) = -2 - 0.5 = -2.5$

$V_1$ (Class 2) $= (-2) + (0.2 \times 0 + 0.8 \times 0) = -2$

$V_1$ (Class 3) $= -2 + (0.6 \times 0 + 0.4 \times 0) = -2$

$V_1$ (Pass) $= .10 + (1 \times 0) = 10$

$V_1$ (Movie) $= 1 + [0.2 \times \overset{-2.5}{0} + 0.4 \times \overset{-2}{0} + 0.4 \times 0\text{-}2]$

$= -1.1$

$V_1$ (Sleep) $= 0.$

## iteration - 2

$V_2$ (Facebook) $= (-1) + [0.9 \times (-1) + 0.1 \times (-2.5)]$

$= -2.15.$

$V_2$ (Class 1) $= (-2) + (0.5 \times (-2) + 0.5\ (-2.15))$

$= -4.075.$

$V_2$ (Class 2) $= (-2) + (0.2 * 0 + 0.8 \times (-2))$

$= -3.6$

$V_2$ (Class-3) $= (-2) + (0.6 \times 10 + 0.4 \times (-1.1))$

$= 3.56.$

$V_2(\text{Pass}) = 10 + 0 = 10$

$V_2(\text{Sleep}) = 0$

$V_2(\text{Movie}) = 1 + [0.2 \times (-4.075) + 0.4 \times (-3.6) + 0.4 \times (3.56)]$

$= 0.169$

again, iteration - 3

$V_3(\text{Facebook}) = -2 + (0.9 \times (-2.15) + 0.1 \times (-4.075)) = -3.3425$

$V_3(\text{Class 1}) = -2 + (0.5 \times (-3.3425) + 0.5 \times (-3.6)) = -5.47125$

$V_3(\text{Class 2}) = -2 + (0.8 \times 3.56 + 0) = 0.848$

$V_3(\text{Class 3}) = -2 + (0.6 \times 10 + 0.4 \times 0.169) = 4.0676$

$V_3(\text{Pass}) = 10$

$V_3(\text{Sleep}) = 0$

$V_3(\text{Movie}) = 1 + (0.2 \times (-5.47125) + 0.4 \times (0.848) + 0.4 \times (4.0676))$

$= 1.87199$

⋮

Hence we go on untill the value function converges and the values for will be after convergence.

$V_{\text{Facebook}} = -23$ $\quad V_{\text{Sleep}} = 0$

$V_{\text{Class 1}} = -13$ $\quad V_{\text{movie}} = 0.8$

$V_{\text{Class 2}} = 1.5$

$V_{\text{Class 3}} = 4.3$

$V_{\text{Pass}} = 10$

or we can have this by solving:

$$V = (I - \gamma P)^{-1} R$$

ow, we have to find optimal value function i.e $v^{*\pi}(s)$ before that we need to varify that the values for the value (state) function converges properly:

For facebook → $V_{Facebook} = -23$

$\Rightarrow V_{Facebook} = -1 + (0.9 \times -23 + 0.1 \times -13)$

$= -1 - \left(\frac{23 \times 9}{10} + \frac{13}{10}\right)$

$= -1 - \left(\frac{207 + 13}{10}\right)$

$= -1 - 22 = -23$ ∴ $V_{Facebook}$ converges (i.e does not change)

again, $V_{class1} = -2 + (0.5 \times 1.5 + 0.5 \times (-23))$

$= -12.75 \approx -13$ (converges)

again, $V_{class2} = -2 + (0.8 \times 4.3 + 0.2 \times 0)$

$= 1.44 \approx 1.5$

$V_{class3} = -2 + (0.6 \times 10 + 0.4 \times 0.8)$

$= 4.32 \approx 4.3$

$V_{movie} = -1 + (0.2 \times (-13) + 1.5 \times 0.4 + 0.4 \times 4.3)$

$= 0.72 \approx 0.8$

and $V_{pass} = 10 + 0 = 10$ and $V_{sleep} = 0$.

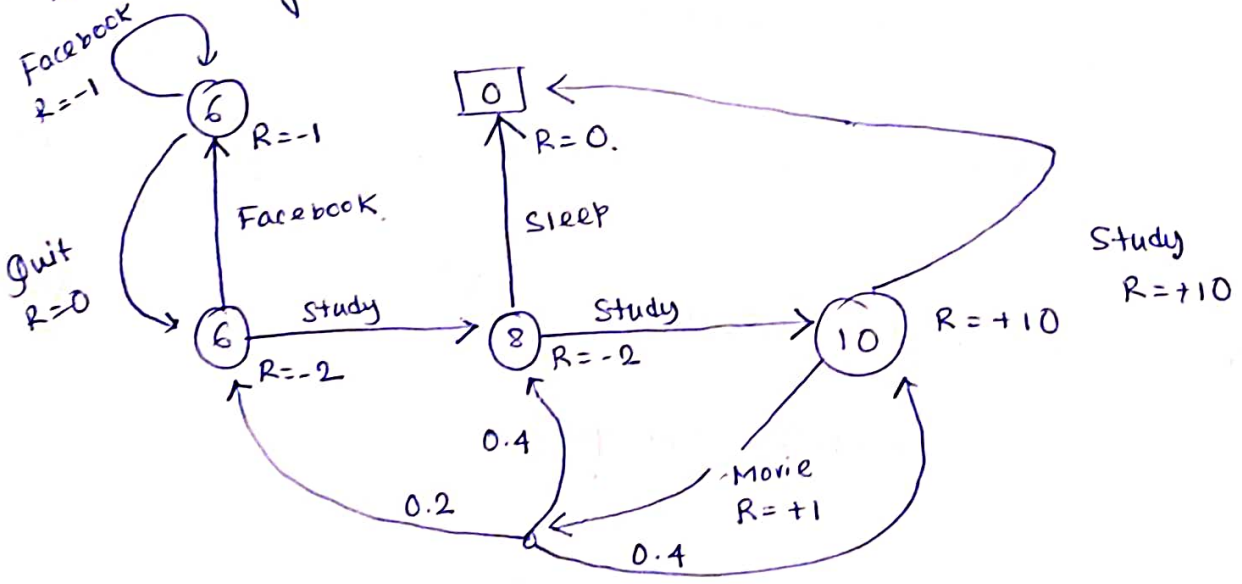
Now, for finding $V^{\pi^*}(s)$ we have to calculate for each state $s \in S$.

$$V^{\pi^*}(s) = V^{\pi}(s) + \gamma \max_{a \in A \text{ or } \pi} \sum_{s' \in S} P_{ss'} V^{\pi}(s')$$

So, as before we initialize all $V(s) = 0 \ \forall \ s \in S$ and calculate

$V^*(s) := R(s) + \max_{\pi} \gamma \sum_{s' \in S} P_{ss'} V(s')$ until $V(s)$ converges to $V^*$.

Now suppose we got :




For Facebook $V^*(\text{Facebook}) = \max_{\pi} V^{\pi}(s)$

$$= \max_{\pi} \{-1+6, 0+6\}$$

$$= \max_{\pi} \{5, 6\}$$

$$= 6$$

$V^*(\text{Class-1}) = \max_{\pi} V^{\pi}(s)$

$$= \max_{\pi} \{-2+8, -1+6\} = \max_{\pi} \{6, 5\}$$

$$= 6$$

Start

$V^*(\text{class-2}) = \max_{\pi} V^{\pi}(s)$

$= \max_{\pi} \{-2+10, 0+0\}$

$= 8$

$V^*(\text{class-3}) = \max_{\pi} V^{\pi}(s)$

$= \max_{\pi} \{10+0, 1+x\}$

$= 10$ (assuming movie's optimal value function is $< 10$).

~~$V^*(\text{movie}) = \max V^{\pi}(s)$~~

~~$= \max$~~
~~10~~

$\therefore \pi^*(\text{Facebook}) = \text{Quit}$

$\pi^*(\text{class 1}) = \text{class 2 (study)}$

$\pi^*(\text{class 2}) = \text{Study (go to class 3)}$

$\pi^*(\text{class 3}) =$ ~~Sleep~~ Study. (then go to sleep)

i.e. Quit Facebook and keep studying!!!

and between movie and study; Study is optimal.

Now, we have to find the optimal Q Function and Corresponding optimal action.

$\therefore Q^*(s,a) = \max_{\pi} \underset{\tau \sim \pi}{E} [R(\tau) | s_0 = s, a_0 = a]$

and $a^*(s) = \arg\max_{a} Q^*(s,a)$

$$\therefore Q^*(s,a) = \sum_{s'} P(s'|a,s)\left(R(s,a,s') + \gamma V^*(s')\right)$$

Now, we have to calculate $Q^*$ for each of the state for each of the actions.

$$\therefore Q^*(\text{Facebook}, \text{Facebook}) = 1 \times (-1 + 6)$$

$$= 1 \times 5 = 5.$$

$$Q^*(\text{Facebook}, \text{Quit}) = 1 \times (0 + 6)$$

$$= 6$$

$$Q^*(\text{class1}, \text{Facebook}) = 1 \times (-1 + 6) = 5$$

$$Q^*(\text{class 1}, \text{Study}) = 1 \times (-2 + 8) = 6$$

$$Q^*(\text{class 2}, \text{Sleep}) = 1 \times (0 + 0) = 0$$

$$Q^*(\text{class 2}, \text{Study}) = 1 \times (-2 + 10) = 8$$

$$Q^*(\text{class 3}, \text{Study}) = 1 \times (10 + 0) = 10$$

$$Q^*(\text{class 3}, \text{Movie}) = 1 \times (1 + x) = 8.4 \text{ (as given in ppt)}$$

where $V^*(\text{movie}) = x$. and $\gamma = 1$ for all the cases here.

Now, for finding $a^*(\text{Facebook}) = \arg\max_a \left(Q^*(\text{FB}, \text{FB}), Q^*(\text{FB}, \text{Quit})\right)$

$$= 6 = \text{Quit}$$

$$a^*(\text{class1}) = \arg\max\left(Q^*(\text{c-1}, \text{FB}), Q^*(\text{c-1}, \text{Study})\right)$$

$$= 6 = \text{Study}$$

$a^*(\text{class 2}) = \text{Study}$ and $a^*(\text{class 3}) = \text{Study}$.