Imperial College London

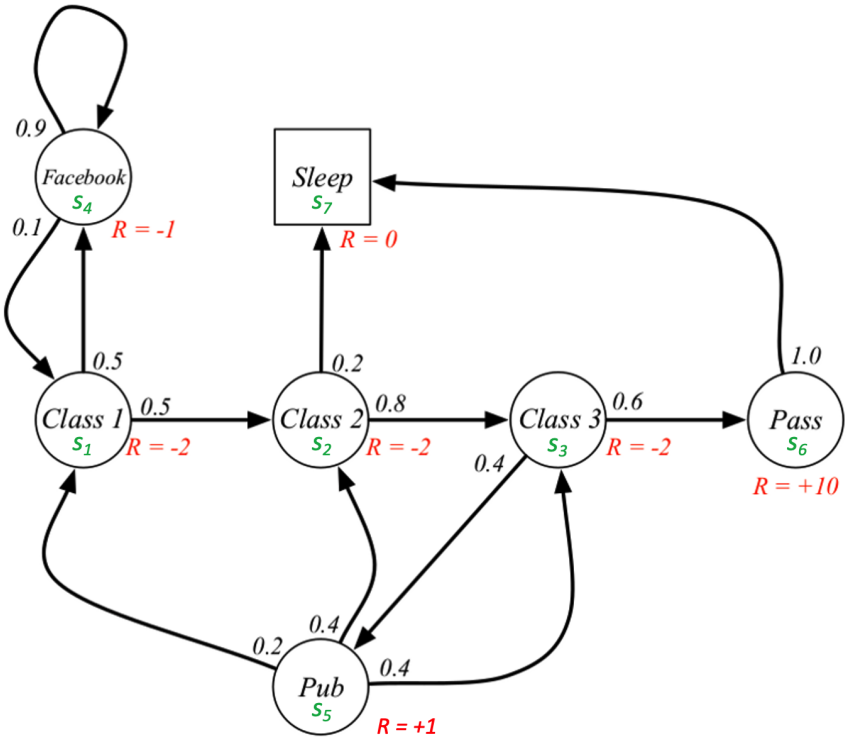
October 11, 2022

Reinforcement Learning – Prof Aldo Faisal Dr Paul Bilokon
Assignment design: Manon Flageat

Lab Assignment 1: Understanding of MRPs and MDPs

Question 1: Simple Markov Reward Process

Consider the Markov Reward Process (MRP) drawn out below (as seen in the lectures), with $\gamma = 1/2$ and with start state $s_1 = s_{class_1}$:



Source: David Silver - UCL lecture

1. What are the terminal states in this MRP?
2. Give an example trace sampled from this MRP, for example:

$$\tau = s_{class_1}\ s_{facebook}\ s_{class_1}\ s_{class_2}\ s_{sleep} = s_1\ s_4\ s_1\ s_2\ s_7$$

3. Give an expression of the state transition probability matrix of this MRP $P_{ss'}$.
4. Give an expression of the reward matrix of this MRP $R_{ss'}$.
5. Consider the following traces sampled from this MRP. What would be the associated returns?

$$\tau = s_{class_1}\ s_{facebook}\ s_{facebook}\ s_{facebook}\ s_{class_1}\ s_{class_2}\ s_{sleep} = s_1\ s_4\ s_4\ s_4\ s_1\ s_2\ s_7$$

$$\tau = s_{class_1}\ s_{class_2}\ s_{class_3}\ s_{pub}\ s_{class_3}\ s_{pass}\ s_{sleep} = s_1\ s_2\ s_3\ s_5\ s_3\ s_6\ s_7$$

Question 2: Understanding Markov Decision Process

An operator wishes to maximise the production of the PhD students' common-room coffee machines. This production is directly related to the machines' condition, and the machines' condition directly relies on their maintenance. We hypothesise that the machines are always used at their full production capacity (which is not totally incorrect in this case).

Here we are considering one of these machines, and we model the problem using a Markov Decision Process (MDP), with the following data:

- The machine has 4 possible states:
  - s_0 = new
  - s_1 = good condition
  - s_2 = poor condition
  - s_3 = broken
- The operator can take 3 possible actions:
  - a_0 = do nothing
  - a_1 = maintain
  - a_2 = renovate
- The machine transition between these states at each time-step, depending on the action taken by the operator, following the transition table given below.
- At each transition (so at each time-step), the machine produce a given quantity of coffee of value-worth p which depends on its state before the transition: $p(s_0) = 100£$, $p(s_1) = 50£$, $p(s_2) = 10£$, $p(s_3) = 0£$.
- Each action has a cost c that needs to be taken into account to compute the reward associated with each (state, action) pair: $c(a_0) = 0£$, $c(a_1) = 30£$, $c(a_2) = 100£$.
- The discount factor is given by $\gamma = 0.6$.

In the questions, we are considering the following deterministic policy π:

- $\pi(s_0) = a_1$ (always maintain new machine)
- $\pi(s_1) = a_0$ (always do nothing when the machine is in good-condition)
- $\pi(s_2) = a_1$ (always maintain machine in bad-condition)
- $\pi(s_3) = a_2$ (always renovate broken machine)

The following table gives the transition probabilities (missing data will be filled later):

Initial state	Action	Final state	Transition probability	Reward
s_0 (new)	a_0 (do nothing)	s_1 (good condition)	5/6	
		s_3 (broken)		
	a_1 (maintain)	s_0 (new)	5/6	
		s_1 (good condition)	1/6	
	a_2 (renovate)	s_0 (new)	1	
s_1 (good condition)	a_0 (do nothing)	s_1 (good condition)	1/6	
		s_2 (poor condition)	4/6	
		s_3 (broken)	1/6	
	a_1 (maintain)	s_1 (good condition)	4/5	
		s_2 (poor condition)	1/5	
	a_2 (renovate)	s_0 (new)	1	
s_2 (poor condition)	a_0 (do nothing)	s_2 (poor condition)	2/3	
		s_3 (broken)	1/3	
	a_1 (maintain)	s_2 (poor condition)		
		s_3 (broken)	1/4	
	a_2 (renovate)	s_0 (new)	1	
s_3 (broken)	a_0 (do nothing)	s_3 (broken)		
	a_1 (maintain)	s_3 (broken)		
	a_2 (renovate)	s_0 (new)	1	

1. Fill in the missing transition probabilities in the transition table.
2. Compute the reward associated with each (state, action) pair, and fill it in the "Reward" column of the transition table.
3. Draw the MDP graph corresponding to policy π (you know the action chose in each state, so you only need to represent the states, the transitions and the rewards).
4. Give an example trace of length 6 sampled from this MDP following the policy π and starting in state s_0.
5. Give the state transition probability matrix of this MDP given the policy π: P^π.
6. Give an expression of the reward matrix of this MDP given the policy π: R_π.
7. Find analytically the value function associated with policy π using these two matrices.
8. Using your previous results, propose a one-step amelioration of your deterministic policy for state s_1 (good condition), $\pi(s_1)$.