Programming Assignment 1

Implementation of the value iteration algorithm for policy evaluation.

Consider a MDP specified as $(\mathcal{S}, \mathcal{A}, \mathbb{P}, \mathcal{R}, \gamma)$. Let $s \in \mathcal{S}$. Then the state value of s for a policy π is

$v^{\pi}(s) = \sum_a \pi(a|s) \sum_r p(r|s,a) + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) v^{\pi}(s').$

To compute $v^{\pi}(s)$ we have the value iteration algorithm, that starts with an estimate $v_0(s), \forall s \in \mathcal{S}$ which is say initialized to 0. Then for every k we define

$v_{k+1}(s) = \sum_a \pi(a|s) \sum_r p(r|s,a) + \gamma \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) v_k(s').$

In this programming assignment you have to implement the above iterative algorithm. Note that in an implementation, we would need to terminate the above iterations after some finite time. We will use the following termination condition.

Stop the iterations if $\max_{s \in \mathcal{S}} |v_{k+1}(s) - v_k(s)|/v_k(s) \leq \epsilon$, where ϵ will be specified. After stopping, assign $v^{\pi}(s) = v_{k_{stop}}(s)$, where k_{stop} is the iteration number at which we have stopped.

1. Write a Python function "valueiteration_pi" that has the inputs - $\mathcal{S}, \mathcal{A}, \mathbb{P}, \mathcal{R}, \gamma, \epsilon$, and π. The function should compute $v^{\pi}(s)$ using the above iterative procedure.
2. Write a Python program that will read the inputs $\mathcal{S}, \mathcal{A}, \mathbb{P}, \mathcal{R}, \gamma, \epsilon$, and π from a file. The file has the following format (line numbers are indicated)
    1. comma separated elements of the state space
    2. comma separated elements of the action space
    3. γ
    4. ϵ
    5. comma separated values of the policy
    6. comma separated values of the reward
    7. comma separated values of the transition probability

    Further specification of these lines follow

    1. comma separated elements of the state space - suppose the state space is the set $\{a, b, c, d\}$, then this line would contain a,b,c,d
    2. comma separated elements of the action space - suppose the action space is the set $\{1, 2\}$, then this line would contain 1,2
    3. γ and ϵ would be single numbers

4. comma separated values of the policy - note that the policy π is in general a random policy, i.e., $\forall s \in \mathcal{S}$ we need to specify $\pi(a|s)$, $\forall a \in \mathcal{A}$. The specification of the pmf is explained using the following example - for the state space $\{a, b, c, d\}$ and action space $\{1, 2\}$, we will give $\pi(1|a), \pi(2|a), \pi(1|b), \pi(2|b), \pi(1|c), \pi(2|c), \pi(1|d), \pi(2|d)$ on line 5. That is, the pmf values would be specified for all actions (in the order specified in the action set) for the first state and then for the second state and so on. The order of the states would be as in the specification of the state space.
5. comma separated values of the reward - we are going to assume that the reward is a deterministic function $r(s, a)$ of the state and action. Again this will specified as a list of comma separated values for all actions (in the order specified in the action set) for the first state, then for the second state and so on. The order of the states would be as in the specification of the state space. For example, for the state space {a,b,c,d} and action space {1,2}, we will give $r(1, a), r(2, a), r(1, b), r(2, b), r(1, c), r(2, c), r(1, d), r(2, d)$ on line 6.
6. comma separated values of the transition probability - the transition probability need to be specified as $p(s'|s, a)$, $\forall s' \in \mathcal{S}$, $\forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$. Again this will be specified as a list of comma separated values of all states (in the order specified in the state space set) for the first action and first state, then for all states for the second action and first state, and so on for all actions for the first state. Then the values for the second state (i.e., s) would be specified. For example, for the state space {a,b,c,d} and action space {1,2}, we will give the following on line 7 $p(a|a, 1), p(b|a, 1), p(c|a, 1), p(d|a, 1), p(a|a, 2), p(b|a, 2), p(c|a, 2), p(d|a, 2),$ $p(a|b, 1), p(b|b, 1), p(c|b, 1), p(d|b, 1), p(a|b, 2) \ldots$.

3. This python program should then call the function that you have written to compute the value function.
4. The value function (which is a sequence of values) computed should be returned from the function.
5. The returned sequence should be printed out to the console as a comma separated list for states, where the states appear in the order specified in the state space. For the example above, you should print out $v^{\pi}(a), v^{\pi}(b), v^{\pi}(c), v^{\pi}(d)$