Introduction to Machine Learning

Spring 2021

University of Science and Technology of China

Lecturer: Jie Wang	Homework 7
Posted: June 5, 2021	Due: June 18, 2021
Name: San Zhang	ID: PBXXXXXXXX

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: 10pts

John and Mary take turns tossing one die; John goes first.

1. The winner is the first player to throw a 6. Who will have an advantage over the other? (Hint: find the one who has a higher probability to win than the other.)

2. We change the rule a little bit. The winner is the first player to throw the same points as that of the previous toss. Who will have an advantage over the other?

Exercise 2: 10pts

Please solve Exercise 1 by reinforcement learning techniques. You can follow the procedure as follows.

1. Model the problem as a finite MDP, i.e., define the

   (a) states $\mathcal{S}$,

   (b) actions $\mathcal{A}$,

   (c) rewards $\mathcal{R}$,

   (d) state transition function (or probabilities),

   (e) reward function (or probabilities),

   (f) policy,

   such that the winning probability corresponds to the value of certain state, and

   (g) verify the Markov property of your MDP.

2. Derive the Bellman equation with respect to the value function.

3. Compute the winning probability by solving Bellman equation.

Exercise 3: Properties of transition matrix 25pts

A matrix is nonnegative (positive) if all its entries are nonnegative (positive). A right (left) stochastic matrix is a square nonnegative matrix with each row (column) adds up to one. Without loss of generality, we study the right stochastic matrix in this exercise. Suppose that $T \in \mathbb{R}^{n\times n}$ is a right stochastic matrix.

1. Show that T has an eigenvalue 1.
2. Let λ be one of T's eigenvalues. Show that $|\lambda| \leq 1$.
3. Show that $I - \gamma T$ is invertible, where $I \in \mathbb{R}^{n\times n}$ is the identity matrix and $\gamma \in (0, 1)$.
4. We now show that $(I - \gamma T)^{-1} = \sum_{i=0}^{\infty} (\gamma T)^i$.

   (a) For $\mathbf{x} \in \mathbb{R}^n$, we define the infinity norm by

   $$\|\mathbf{x}\|_\infty = \max_i |x_i|.$$

   The induced norm of matrix $M \in \mathbb{R}^{m\times n}$ is

   $$\|M\|_\infty = \max_{\|\mathbf{x}\|_\infty \leq 1} \|M\mathbf{x}\|_\infty.$$

   i. Show that $\|M\|_\infty = \max_i \sum_{j=1}^{n} |m_{i,j}|$.
   ii. Show that $\|cM\|_\infty = |c|\|M\|_\infty$ for any $c \in \mathbb{R}$.
   iii. Show that $\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty$ holds for any $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$.

   (b) Show that the sequence $I, \gamma T, \sum_{i=0}^{2} (\gamma T)^i, \ldots, \sum_{i=0}^{n} (\gamma T)^i, \ldots$ is Cauchy. (Hint: a matrix sequence $\{A_p\}$ is Cauchy in $(\mathbb{R}^{m\times n}, \|\cdot\|_\infty)$, if given any $\epsilon > 0$, there is an integer $N \geq 1$ such that $\|A_p - A_q\|_\infty < \epsilon$, whenever $p, q \geq N$.)

   (c) Combining the result in the last part and the fact that $\mathbb{R}^{n\times n}$ is complete, we can conclude that $\sum_{i=0}^{\infty} (\gamma T)^i$ converges to a matrix which we denote by L. Show that $(I - \gamma T)^{-1} = L$. (Hint: you need to show that $(I - \gamma T)L = \lim_{n\to\infty} \sum_{i=0}^{n} (I - \gamma T)(\gamma T)^i$.)

Exercise 4: Grid World with a Given Policy 15pts

Consider the grid world shown in Figure 1. The finite state space is $\mathcal{S} = \{s_i : i = 1, 2, \ldots, 11\}$ and the finite action space is $\mathcal{A} = \{\text{up, down, left, right}\}$.
State transition probabilities After the agent picks and performs a certain action, there are four possibilities for the next state: the destination state, the current state, the states to the right and left of the current state. If the states are reachable, the corresponding probabilities are 0.7, 0.1, 0.05, and 0.15, respectively; otherwise, the agent stays where it is. The game will terminate if the agent arrives at s_{10} (loss) or s_{11} (win).
Reward After the agent picks and performs a certain action at its current state, it receives rewards of 100, -100, and 0, if it arrives at states s_{11}, s_{10}, and all the other states, respectively.
Policy In Figure 1, the arrows show the policy $\pi : \mathcal{S} \to \mathcal{A}$ for the agent. The random variable S_t is the state at time t under the policy π.

1. Find the matrix $M \in \mathbb{R}^{11\times 11}$ with $m_{i,j} = \mathbf{P}(S_{t+1} = s_i | S_t = s_j)$, i.e., the conditional probability of the agent moving from s_j to s_i .

2. Suppose that the initial state distribution is uniform distribution, that is $\mathbf{P}(S_0 = s_i) = 1/11$, $i = 1, \ldots, 11$.

   (a) Find the distributions $\mathbf{P}(S_1)$ and $\mathbf{P}(S_2)$ by following the policy π.

   (b) Show that the agent would finally arrive at either s_{10} or s_{11}, i.e.,

   $$\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 9.$$

   (c) Please find $\lim_{t\to\infty} \mathbf{P}(S_t = s_{10})$ and $\lim_{t\to\infty} \mathbf{P}(S_t = s_{11})$.

3. Find the value function corresponding to π, where the discount factor $\gamma = 0.9$.

4. Show that the result in (2b) holds for any initial probabilities we choose for $\mathbf{P}(S_0 = s_i)$, $i = 1, \ldots, 11$.

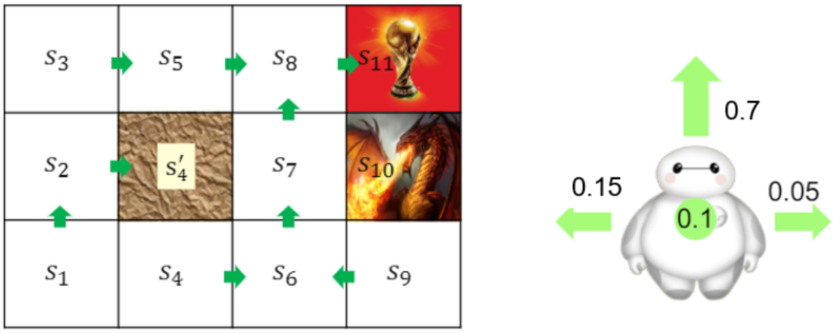

Figure 1: Illustration of a grid world with a policy.

Exercise 5: Optimal Policy 10pts

Consider the grid world problem described in Exercise 4. Let π^* be the optimal policy, V^* the corresonding value function, Q^* the corresponding Q function, and $\gamma = 0.9$.

1. Please derive the Bellman Equation in terms of the value function V^* and the Q function Q^*, respectively.

2. Please choose one of the algorithms we introduced in class to find π^* and V^* respectively and write their pseudocode (hand in your code if you have one).

3. Please design a reward scheme such that following the resulting optimal policy will never lose. Specifically, you need to derive the resulting optimal policy (the proof is not required) and show

$$\lim_{t \to \infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 10,$$

   whenever $\mathbf{P}(S_0 = s_{10}) = \mathbf{P}(S_0 = s_{11}) = 0$.

Exercise 6: Value Iteration 25pts

Consider a Markov Decision Process with bounded rewards and finite state-action pairs. The transition probability is $\mathbf{P}[s'|s,a]$, the discounted factor is $\gamma \in (0,1)$, and the reward function is $r(s,a)$. Let $\pi : \mathcal{S} \to \mathcal{A}$ be a deterministic policy.

1. (**The Contraction Mapping Theorem**) Given the normed vector space $(\mathbb{R}^n, \|\cdot\|_\infty)$, a map $f : \mathbb{R}^n \to \mathbb{R}^n$ is called a contraction, if there is some constant $\alpha \in [0,1)$, such that $\|f(x) - f(y)\|_\infty \leq \alpha\|x - y\|_\infty$, for all $x, y \in \mathbb{R}^n$. Moreover, x is the fixed point of the function $f : \mathbb{R}^n \to \mathbb{R}^n$ if $f(x) = x$.
   Given the contraction $f : \mathbb{R}^n \to \mathbb{R}^n$,

   (a) start with an arbitrary point $x_0 \in \mathbb{R}^n$, and define a sequence $x_n = f(x_{n-1})$. Please show that the sequence $(x_n)_n$ converges to the fixed point of f.

   (b) please show that f admits a unique fixed point x.

2. Let V^k denote the value function after the k^{th} iteration of the value iteration algorithm. Please show that value iteration achieves linear convergence rate, that is

$$\|V^* - V^{k+1}\|_\infty \leq \gamma\|V^* - V^k\|_\infty.$$

Exercise 7: Policy Iteration 15pts

Consider the problem described in Exercise 6. Let $Q^\pi(s,a)$ be the accumulated reward by performing the action a first and then following the policy π.

1. Find the Bellman Equation for Q^π.
2. Consider a new policy π' given by

   $$\pi'(s) = \mathop{\mathbf{argmax}}_{a\in\mathcal{A}} Q^\pi(s,a).$$

   Note that if $\mathbf{argmax}_{a\in\mathcal{A}} Q^\pi(s,a)$ is not unique, we can choose one action arbitrarily. Show that $V^{\pi'}(s) \geq V^\pi(s)$ for all $s \in \mathcal{S}$.

Exercise 8: Q learning algorithm for deterministic MDP 15pts

Consider the Q learning algorithm for any deterministic MDP with non-negative rewards and finite state-action pairs. Assume that we initialize all $\hat{Q}$ values to zero. Let $\hat{Q}_n(s,a)$ denote the learned $\hat{Q}(s,a)$ value after the n^{th} iteration of the training procedure in Q learning algorithm.

1. Please show that $\hat{Q}$ values never decrease during training, that is

$$\hat{Q}_{n+1}(s,a) = r + \gamma \max_{a'} \hat{Q}_n(s',a') \geq \hat{Q}_n(s,a),\ \forall s,a,n.$$

2. Please show that throughout the training process, every $\hat{Q}$ value will remain in the interval between zero and the optimal Q function Q^*, that is

$$0 \leq \hat{Q}_n(s,a) \leq Q^*(s,a),\ \forall s,a,n.$$