ECE 276 Assignment 2 : Tabular Methods

October 16, 2019

In this assignment, we will solve a simple grid world problem called 'FrozenLake-v0' in OpenAI gym using both model based and model free methods. To learn how to set up the environment and interact with it take a look at the OpenAI website.(More about the environment can be found on the OpenAI github page)
Note: Use the virtual environment from Assignment 1.

Question 1 - Model based methods

1. Describe the environment state and action spaces, and reward function. Given a state and an action, is the state transition deterministic?

2. Given a Markov Decision Process described by $\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma$, where $\mathcal{S} \in \mathbb{R}^n$ is the state-space, $\mathcal{A} \in \mathbb{R}^m$ is the action space, $\mathcal{R} : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the reward function, $P : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \to [0, 1]$ is the transition probability and γ is the discount factor. Show that for a deterministic policy $\pi(s)$, the value function $v(s)$ can be expressed as:

$$v(s) = \sum_{s' \in \mathcal{S}} p(s'|s, a)[r(s, a, s') + \gamma v(s')] \tag{1}$$

   where $p(s'|s, a) \in \mathcal{P}$ and $r(s, a, s') \in \mathcal{R}$. Assume that the state and action spaces are discrete.

3. Write a function `TestPolicy(policy)`, that returns the average rate of successful episodes over 100 trials for a deterministic policy. What is the success rate of a policy given by $\pi(s) = (s + 1)\%4$, where % is the modulus operator.

4. Write a function `LearnModel`, that returns the transition probabilities $p(s'|a, s)$ and reward function $r(s, a, s')$. Estimate these values over 10^5 random samples.

5. Write a function `PolicyEval` for evaluating a given deterministic policy and with the help of this function implement a policy iteration method to solve this environment over 50 iterations. Plot the average rate of success of the learned policy at every iteration.

6. Write a function `ValueIter` that returns a deterministic policy learned through value-iteration over 50 iterations. Plot the average rate of success of the learned policy at every iteration.

Question 2 - Model free methods

1. Solve the environment using Q-learning over 5000 episodes. For exploration during training, take random actions with probability 1-e/5000 where e is the number of current episode. Plot the success rate of the learned policy at an interval of 100 episodes.

   (a) Train the policy using the following learning rates with $\gamma = 0.99$.Report what you observe.

   $$\alpha \in \{0.05, 0.1, 0.25, 0.5\}$$

   (b) Train the policy using the following discount factors with $\alpha = 0.05$. Report what you observe.

   $$\gamma \in \{0.9, 0.95, 0.99\}$$

2. In the previous question, the exploration was linearly annealed. Solve the environment using Q-learning by proposing a different strategy to explore. Find a suitable α and γ for your method. Report your strategy and training results.