CS 5751 – Spring 2018 – Homework 2
Assigned: 01/30/2018
Due: 02/13/2018
Total points: 100 pts.
Submit a soft copy to canvas. Remember to write your name at the top of each file you submit.

Objectives: The objectives of this homework are the following:

- Explore a solution to the k-armed bandit problem.
- Verify that the online algorithm for calculating the mean works.
- Practice thinking about the elements of reinforcement learning.
- Learn how to do policy prediction by solving the associated system of equations.

Notes:

- This homework is to be done individually. You may discuss with your classmates, but the work that you write must be your own.
- If you choose Python, then you need to make sure that you have the following packages: numpy, jupyter. See the slides posted on canvas to learn how to install Python and R. If you are using Linux or Mac, you can install these packages by typing the following at the command prompt: pip install numpy, jupyter.

Activity 1: (20pts.) (k-armed Bandit) Perform activities b and c:

a) (0 pts.) Read Sections 2.1 to 2.5 from the Reinforcement Learning book (January 2018 version). Then, download from the website https://github.com/ShangtongZhang/reinforcement-learning-an-introduction the code to replicate Figure 2.2. Convert this code into a Jupyter notebook. Try to use different cells. For this, use your judgement: this does not mean having each instruction in a separate cell. Just try to group them in a meaningful way. Then run the code and replicate Figure 2.2 of the book.
b) (10 pts.) Now, modify this code to implement exercise 2.5 from the book. Then generate the resulting figures.
c) (10 pts.) What do you observe in your experiments results of part b)? What made this problem different?

Write a Jupyter notebook named yourLastName_hw2_q1.ipynb with the answers to this activity.

Activity 2: (25pts.) (Online algorithm for the mean) Write a Jupyter notebook named yourLastName_hw2_q2.ipynb in either Python or R to do the following tasks:

a) (5 pts.) Download the Air quality dataset from the UCI dataset repository web site https://archive.ics.uci.edu/ml/datasets/Air+Quality and load it into a data frame. Be careful because the attribute separator is not the comma, but the semicolon, so the default parameters for read_csv won't work.
b) (15 pts.) Write an online algorithm to compute the mean of a sequence of n numbers $\{A_i : 1 \leq i \leq n\}$. This algorithm should start with $A_0 = 0$, and then as num-

bers come (that is, as *i* increases), it should update the mean of the numbers it has seen so far.

c) (5 pts.) Run your algorithm on the temperature attribute of the dataset, and plot $mean(A_i)$ vs. i. Then check that $mean(A_n)$ is indeed the mean of the all the points $\{A_i : 1 \leq i \leq n\}$. To verify this, you can use the *mean* function of R or numpy.

Activity 3: (25pts.) (Reinforcement Learning) Read Section 3.1 of the RL book (January 2018 draft), then devise three example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the three examples as different from each other as possible. The framework is abstract and flexible and can be applied in many different ways. Stretch its limits in some way in at least one of your examples.

Write a PDF file yourLastName_hw2_q3.pdf with the answers to this activity.

Activity 4: (25pts.) (Policy prediction for Gridworld) (We'll see this topic on Thursday 02/01) Assume that your task is the one described in Figure 3.2 in the reinforcement learning book, where you have:

- A 5×5 gridworld.
- Your agent's available actions at each state are: going up, down, right or left.
- The policy π that your agent is following chooses these four actions with equal probability.
- The actions are deterministic.
- Running into the edges of the grid would produce a reward of -1 and leave you at the same state.
- Moving from one state to another produces a reward of 0, except that if you are in state A, then any action will get your agent to state A' with a reward of 10, and if you are in state B, any action will get your agent to state B' with a reward of 5.

Now, write a Jupyter notebook in R or Python named yourLastName_hw2_q4.ipynb with the answers to this activity:

a) Write down the Bellman equations for $v_\pi(s)$ for the states (0,0), (0,1) and (2,1) in this Gridworld, using the policy described above. This means plugging in the right numbers and probabilities, not just writing the abstract formulas.

b) Using a solver for systems of linear equations, solve the complete system of value-state Bellman equations to obtain $v_\pi(s)$.

c) Using the solution you obtained in part 5b, print an array such that at state *s* it displays the value $v_\pi(s)$, for the policy π. In other words, this exercise is asking you to replicate Figure 3.2 in the book.