CMPSCI 687: Course Project
Fall 2020

Philip S. Thomas
University of Massachusetts
pthomas@cs.umass.edu

For this project, we will simulate a high-risk real-world application. You are faced with a decision-making problem (you might envision optimizing diabetes treatments or optimizing the decision of which ad to show a specific user on a web page). There is a currently deployed policy, which was used to collect data. We will provide you with this data, and your job is to provide us with a policy that is better than our current policy.

While the true state of the world is unknown to us, we have taken our observations and clustered them into 18 possible values, which you might think of as a discrete state representation with 18 possible discrete values, $\{0, 1, \ldots, 17\}$. There are four actions available, $\{0, 1, 2, 3\}$. We have selected $\gamma = 0.95$, and know that all rewards will be in the closed interval $[0, 10]$. We do not know anything about the horizon, and do not have access to the parametric form of the currently deployed policy. It is possible that a state that can occur does not occur within the data that we have collected. The data that we have collected is in a text file that you can find here (warning: the .zip file is 68MB, but the unzipped file data.csv is 750MB).

The first line of this file is the number of episodes. Each episode is then represented by 1) one line that indicates the number of time steps in the episode, 2) one line per time step of the episode. The latter lines are of the form $S_t, A_t, R_t, \pi_b(S_t, A_t)$. Notice that, even though we write S_t here, the provided "states" are actually our state representation, which may not be Markovian. For example, the following would correspond to two episodes of data, the first with one time step and the second with two:

```
2
1
7,2,10,0.5
2
0,1,0,0.3
17,0,1,0.1
```

You will submit 100 policies as text files `policy1.txt`, `policy2.txt`, ..., `policy100.txt`. Each of these text files will contain 18×4 real numbers, with each number on its own line (there should be no commas in your submitted .txt files). When in state s, the probability of action a is

$$\pi(s, a) = \frac{e^{\theta_{4s+a}}}{\sum_{a'} e^{\theta_{4s+a'}}},$$

where θ_i is the i^{th} number in the policy file. Note: You may use any policy representation during learning, as long as you convert it to this form for submission.

For this project, you may use any programming language and any software libraries that you find online that were published prior to November 13, 2020. You may discuss the project with other students. However, you are not allowed to collaborate with other students when coding—your code must be your own. You may implement any algorithm—deciding what methods to use is part of the project.

1 Due Date

Submissions are due December 11, 2020 at 11:59pm Amherst-time.

2 Grading

Your grade will be the percent of the policies that you submit that produce an expected discounted return of at least 1.41537. The policies that you submit do not need to be unique—you may submit the same policy 100 times (which would result in your grade being either 100% or 0%).

3 Submission

You should submit a .zip file containing 100 text files and a directory. The 100 text files are the submitted policies, and should be name `policyX.txt` for $X \in \{1, 2, \ldots, 100\}$. The directory should be named "source", and should include your code and instructions for compiling and running your code in a file called `readme.txt` within the root of the directory. There is no required structure/format for your source code within the "source" directory other than the requirement for `readme.txt` to contain compilation/running instructions and for it to be in the root of the source directory. You must ensure that running your code produces the provided policy files—submitting code that does not produce the policy files you submitted will be considered an academic honesty violation.

4 Project Changes

If any changes are made to the project after it is posted, this document will be updated and a description of changes included below.