COMP-579: Reinforcement Learning - Assignment 3

Posted Tuesday, March 21, 2023
Due Tuesday, April 4, 2023

The assignment can be carried out individually or in teams of two.

1. **Offline RL [100 points]**

   For this assignment, you will start from your preferred agent that you coded last time for the cart-pole domain from the Gym environment suite:

   `https://gymnasium.farama.org/environments/classic_control/cart_pole/`

   You will use this agent to gather 500 behavior episodes in the task. You should also gather 500 episodes using a uniformly random policy. We will now compare two approaches:

   - Simple imitation learning: for this approach, you will use logistic regression to imitate the action observed in each state
   - Fitted Q-learning, which is a precursor to some of the algorithms discussed in class (this is basically using Q-learning targets, but only on the batch of data given, and K is a hyperparameter):

     1. collect dataset $\{(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_i)\}$ using some policy
     2. set $\mathbf{y}_i \leftarrow r(\mathbf{s}_i, \mathbf{a}_i) + \gamma \max_{\mathbf{a}'_i} Q_\phi(\mathbf{s}'_i, \mathbf{a}'_i)$
     3. set $\phi \leftarrow \arg\min_\phi \frac{1}{2} \sum_i \|Q_\phi(\mathbf{s}_i, \mathbf{a}_i) - \mathbf{y}_i\|^2$

     $K\times$

   For this experiment, you should use the same features as last time: discretize the state variables into 10 bins each; weights for the Q-function start initialized randomly between -0.001 and 0.001. Like last time, use 2 learning rate settings. However, this time there will be no exploration, as we are just using the collected data, and we will also only perform one run.

   You will have to create several datasets to test the approach, of sizes: 100 episodes, 250 episodes (obtained by adding 150 episodes to the previous ones) and 500 episodes. There are 3 conditions for this data: (a) All episodes are from the "expert" policy (b) All episodes are from the random policy (c) you select at random half of the episodes form the expert policy and half from the random policy. This will give you 9 datasets on which to train.

   Once you have trained your two estimators to completion, run the greedy policy obtained for 100 episodes and record the returns. Plot a bar chart, showing the average and standard error of the recorded returns for each algorithm and each condition, in each of the data set sizes. (for each data set size, you will have 6 bars). Also draw two horizontal lines, indicating the average return of the expert and of the random policy (evaluated for the same number of episodes)

   Write a small report that describes your experiment, including how you decided to stop the training, the results, and the conclusions you draw from this experimentation. Comment on whether the algorithms are matching and/or exceeding the performance of the policies used to generate the data. Comment also on the impact of the data set size and data quality on the results.

   Extra credit 20 points: Carry out the same work with a multi-layer perceptron as the function approximator