

Responsible Machine Learning

Assignment 4

10 points

©Patrick Hall 2021

June 8, 2022

In Assignment 4, you will work with your group to “red-team” your best model following the instructions below and treating your best model as a black box. A [template](#) has been provided with examples of simple model extraction and adversarial example attacks. For those of you who use Python virtual environments, a basic [requirements.txt](#) file is also available for the template.

Please let me know immediately if you find typos or mistakes in this assignment or related materials.

1 Conduct a white-hat model extraction attack.

Cells 10–16 demonstrate a simple, but effective, model extraction attack. My example model extraction attack uses a single decision tree, that I then plot and use to craft adversarial examples. I’d like for you to try a decision tree extraction attack, but you don’t have to use my code. I imagine getting `graphviz` installed could be difficult for some, so feel free to use your favorite kind of decision tree if the template code proves difficult to run. (Basic instructions for installing `graphviz` are available at the bottom of the [class website](#).)

You may call `predict()` on your best model only one time to perform the extraction attack.

2 Find adversarial examples for your model.

Cells 17–20 use the information from the extracted decision tree to craft highly effective adversarial examples. You must create at least one adversarial example (row of data) that can evoke low predictions from your best model, and at least one adversarial example that can evoke high predictions.

You may call `predict()` on your best model only one time to test your adversarial examples.

3 Submit Code Results.

Your deliverable for this assignment is to update your group’s GitHub repository to reflect this “red-teaming” exercise. The model extraction attack is worth 5 points, and the adversarial examples are worth another 5 points.

In the real world, after performing this analysis, you would want to contact your manager and your organization’s IT security team to discuss the vulnerability. You would argue to put in place authentication on the vulnerable model end point and also argue that monitoring the model’s production scoring queue for random data and training data would be advisable, if possible.

Your deliverables are due Saturday, June 18th, at 11:00 AM ET.

Note that you may also improve Assignment 1 or 3 scores throughout the Summer I Session to improve your ranking, your Assignment 1 grade, your Assignment 3 grade, and your final project grade. Moving forward, you'll need to be able to show that your new predictions preserve $\text{AIR} > 0.8$ for all protected groups.