Automation of Biological Research: 02-450/02-750

Carnegie Mellon University

Homework 4

Version 1.0; updated 4/4/2021

Due: May 7, 2021 by 11:59pm

Overview

This is your final assignment for the semester. It has two interrelated parts, each worth 100 points towards your final grade. As with previous assignments, you are allowed to work in teams of up to 3 people. Of course, you may also work alone, if you prefer.

You will hand in a single PDF file to Canvas that contains the following information:

- The names and andrew ids of the team members.
- A link to an mp4 file of a video presentation that you will create (see below)
- A link to a pdf of the slides you will create (see below)
- A link to a zip (or gz) file with any code that you write (see below). The code should be able to reproduce any figures you make.

Question 1 (100 points)

You will read and prepare a short presentation (~15 – 20 minutes) on the following paper:

- Large scale active-learning-guided exploration for in vitro protein production optimization, Borkowski, *et al*, *Nature Communications* 11(1872), 2020

Imagine that you were going to give a short lecture on this paper to the rest of the class. That is, your audience knows the core concepts and algorithms of Active Learning and Sequential Model Based Optimization, but they are not familiar with this specific paper. Create a video that explains the paper to your classmates. You are free to structure the presentation as you see fit, but it should cover each of the following topics:

- The scientific context. What is cell-free protein synthesis? What advantage(s) does it have over traditional approaches to protein synthesis?
- The problem specification. What are they trying to optimize? What are the design parameters?

- The algorithmic strategy. Which base learner(s), data access model, and query selection strategy did they use?
- The experiments and the results.

You may use figures from the paper and the supplementary information in your slides, but any explanatory text should be your own.

Grading:

Your grade for question 1 will be based on your ability to explain the paper (80 points), and the quality of your presentation (20 points).

Question 2 (100 points)

The file *DataPool.csv* contains the 1,017 designs that were evaluated in the paper you read for Question 1. Each design is an 11-dimensional vector, were each component is one of the 11 adjustable parameters. The final column of each row is the measured yield.

In this question, you will select and run **two algorithms** on the data in *DataPool.csv*. The first algorithm should optimize yield (like the paper). The second algorithm should try to maximize the *accuracy* of the regression model (i.e., not necessarily maximize yield).

Experiment budget: The budget in the paper was 1,017 designs. Your budget is 510.

Batch size: The batch size in the paper was 102 designs per batch. You are free to select any batch size you like.

Algorithms & Software: You are also free to use any algorithm or software you wish, including *modAL*. If you prefer, you can design, implement, and evaluate your own algorithm.

Create a short video (~10 minutes), as if you were presenting the results of your experiments to the rest of the class. You can assume that they are familiar with the problem and the data set, so you can focus your discussion on the methods you selected, and the results you obtained. You may want to combine the video from questions 1 & 2 (i.e., a single 25-30 minute presentation). However, if you prefer, you can create separate videos for each question. Either way, you are free to structure the presentation as you see fit, but it should cover each of the following topics:

- The algorithmic strategy for scenario 1 (i.e., the optimization problem). Which base learner and query selection strategy did you use?

- The algorithmic strategy for scenario 2 (i.e., maximizing accuracy). Which base learner and query selection strategy did you use?
- The experiments and the results.
  - What batch size(s) did you use?
  - Create plots that show the progress of the algorithms, as a function of the number of experiments/rounds.
    - Recall that your budget is 510 experiments, so you can use the remaining 507 rows in *DataPool.csv* as a test set, when computing accuracies.
  - Compare the test errors of the model produced under each scenario.
  - The median yield and correlation (R^2) between predictions and actual yields after **round 5** in the paper (i.e., after ~510 experiments) were ~10 and ~0.6, respectively (see Figure 3). How do your results compare?

Grading:

Your grade for question 2 will be based on your ability to explain your method and the results (50 points), whether the methods you selected are appropriate (20 points), your code (20 points), and the quality of your presentation (10 points). Your grade **will not** be based on the actual yield or correlations you obtain.