Machine Learning - Problem Set 2

PPHA 30545 - Professor Clapp
Winter 2023

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Wednesday, February 1st.** You are welcome (and encouraged!) to form study groups (of no more than 2 students) to work on the problem sets and mini-projects together. But you must write your own code and your own solutions. Please be sure to include the names of those in your group on your submission.

You should submit your answers in one of two ways:

1. As a single PDF containing BOTH a write-up of your solutions that directly integrates any relevant supporting output from your code (e.g., estimates, tables, figures) AND your code appended to the end of your write up. You may type your answers or write them out by hand and scan them (as long as they are legible). Your original code may be a Python (*.py) or Jupyter Notebook (*.ipynb) file converted to PDF format. OR

2. As a single PDF of a Jupyter Notebook (*.ipynb) file with your your solutions and explanations written in Markdown.[1]

Regardless of how you submit your answers, be sure to make it clear what question you are answering by labeling the sections of your write up well. Also, be sure that it is immediately obvious what supporting output from your code (e.g., estimates, tables, figures) you are referring to in your answers. In addition, your answers should be direct and concise. Points will be taken off for including extraneous information, even if part of your answer is correct. You may use bullet points if they are beneficial. Finally, for your code, please also be sure to practice the good coding practices you learned in PPHA 30535/6 and comment your code, cite any sources you consult, etc.

You are allowed to consult the textbook authors' websites, Python documentation, and websites like StackOverflow for general coding questions. You are not allowed to consult material from other classes (e.g., old problem sets, exams, answer keys) or websites that post solutions to the textbook questions.

1. Do the following questions from Chapter 4 of the *Introduction to Statistical Learning* textbook:

   (a) Question 5

   (b) Question 6

   (c) Question 7

   (d) Question 14, questions (a) - (g)

[1] Converting a Jupyter Notebook to PDF is not always straightforward (e.g., some methods don't wrap text properly). Please ensure that your PDF is legible! We will deduct points if we cannot read your PDF (even if you have the correct answers in your Notebook).

- Note: In part (c), please use an 80/20 training/test split (80% of the observations in the training set and 20% in the test set). To avoid confusion among partners and facilitate grading, please also set `random_state=42` when you split the data. This controls how the data are shuffled (randomly ordered) before the split is done. If you're curious, you can try different values to see how it affects your results.

2. Do the following questions from Chapter 5 of the *Introduction to Statistical Learning* textbook:

   (a) Question 5

   - Note: In parts (b) and (d), please use a 70/30 training/validation set split and set `random_state=42`.
   - In part (c), keep the 70/30 training/validation set split, but set `random_state` equal to 2, 6, and 9 to obtain the three different splits of the observations.