Machine Learning for Public Policy - Problem Set 1

The University of Chicago - Harris School of Public Policy
PPHA 30545 - Professors Clapp and Levy
Winter 2025

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Thursday, January 23rd.** There will be separate Gradescope assignments for R and Python students. Please be sure to submit to the version that matches the coding language of the lab section you are enrolled in.

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 format your submission 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 R (*.rmd) or Python (*.py) file converted to PDF format. OR

2. As a single PDF of an R Markdown (*.rmd), Jupyter Notebook (*.ipynb), or Quarto (*.qmd) document with your your solutions and explanations written in Markdown.[1]

Regardless of how you format your answers, be sure to make it clear what question you are answering by labeling the sections of your write up well and assigning your answers to the appropriate question in Gradescope. 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 Data and Programming and comment your code, cite any sources you consult, etc.

You are allowed to consult the textbook authors' website, Python documentation, and websites like StackOverflow for general coding questions. If you get help from a large language model (LLM) or other AI tool (e.g., ChatGPT), you must provide in the query string you used and an explanation of how you used the AI tool's response as part of your answer. 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] 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).

1. (ISL: Chapter 2, Question 3)[2] We now revisit the bias-variance decomposition.

   (a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.[3]

   (b) Explain why each of the five curves has the shape displayed in Question (1a).

2. (ISL: Chapter 2, Question 5) What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

3. (ISL: Chapter 2, Question 10) This exercise involves the `Boston` housing data set.[4]

   (a) To begin, load in the `Boston` data set, which is available on Canvas and/or can be loaded as part of the `ISLR2` library (in R) or `ISLP` library (in Python).

   (b) How many rows are in this data set? How many columns? What do the rows and columns represent?

   (c) Make some pairwise scatterplots of the predictors (columns) in this data set (and display a minimum of five of those plots). Describe your findings.

   (d) Are any of the predictors associated with per capita crime rate? If so, explain the relationship (for the predictors whose plots you displayed in the previous question).

   (e) Do any of the Boston census tracts appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

   (f) How many of the census tracts in this data set bound the Charles River?

   (g) What is the median pupil-teacher ratio among the census tracts in this data set?

   (h) Which Boston census tract has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

   (i) In this data set, how many of the census tract average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

4. (ISL: Chapter 3, Question 3) Suppose we have a data set with five predictors, X_1 =GPA, X_2 =IQ, X_3 =Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Level. The response is starting salary after

[2] Problem set questions are taken from the *Introduction to Statistical Learning* (ISL) textbook. I'll note the corresponding textbook question for your reference, but please be aware that the problem set questions may be modified for clarity or pedagogical reasons.

[3] You can sketch the curves by hand. You don't need to plot the curves in Python or R (although you're welcome to if you'd like to).

[4] Each observation in the `Boston` housing data set is a different census tract.

graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, and $\hat{\beta}_5 = -10$.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 GPA + \hat{\beta}_2 IQ + \hat{\beta}_3 Level + \hat{\beta}_4 (GPA \times IQ) + \hat{\beta}_5 (GPA \times Level)$$

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates.

ii. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates.

iii. For a fixed value of IQ and GPA, high school graduates earn more, on average, than college graduates provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, college graduates earn more, on average, than high school graduates provided that the GPA is high enough.

(b) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

5. (ISL: Chapter 3, Question 15) This problem involves the `Boston` data set (which is used as an example in the lab at the end of Chapter 3 in the textbook). We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response (outcome), and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots (and display a minimum of five of those plots) to back up your assertions. [5]

(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

(c) How do your results from Question (5a) compare to your results from Question (5b)? Create a plot displaying the univariate regression coefficients from Question (5a) on the x-axis, and the multiple regression coefficients from Question (5b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X (that isn't an indicator variable), fit a model of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.[6]

[5] Questions (5a) and (5d) ask you to plot/run a separate regression for each predictor. Plotting or running each regression individually is an acceptable approach, but its tedious/repetitive. Looping over the predictors and/or using functions, then displaying a summary of the results is a better approach.

[6] Please summarize the relevant estimates from your regressions in a table rather than presenting the full output from every regression you run.