Machine Learning for Public Policy - Problem Set 2

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, February 6th.** 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, R/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 4, Question 5)[2] We now examine the differences between LDA and QDA.

   (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? Briefly explain.

   (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set? Briefly explain.

   (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

   (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

2. (ISL: Chapter 4, Question 6) Suppose we collect data for a group of students in a statistics class with variables X_1 =hours studied, X_2 =undergrad GPA, and Y =receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.

   (a) Predict the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in (the previous question) need to study to have a 50% chance of getting an A in the class?

3. (ISL: Chapter 4, Question 7) Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.[3]

4. (ISL: Chapter 4, Question 14, parts (a) - (g)) In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

   (a) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. Note you may find it helpful to add a column `mpg01` to the data frame.

   (b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

[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] Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$. You will need to use Bayes' theorem.

(c) Split the data into a training set and a test set. Please use a 50/50 training/test split (50% of the observations in the training set and 50% in the test set). To avoid confusion among partners and facilitate grading, Python students should set `random_state=22` and R students should set `set.seed(22)` when splitting the data.[4]

(d) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in Question (4b). What is the test error of the model obtained?

(e) Perform QDA on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in Question (4b). What is the test error of the model obtained?

(f) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with mpg01 in Question (4b). What is the test error of the model obtained?

(g) Perform naive Bayes on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in Question (4b). What is the test error of the model obtained?

5. (ISL: Chapter 5, Question 5) In Chapter 4, we used logistic regression to predict the probability of `default` using `income` and `balance` on the `Default` data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) Fit a logistic regression model that uses `income` and `balance` to predict `default`.

   (b) Using the validation set approach, estimate the test error of this model. In order to do this, you should perform the following steps:

      i. Split the sample set into a training set and a validation set. In this question, please use a 70/30 training/validation set split. Please set `random_state=42` (Python) or set `set.seed(42)` (R) when splitting the data.

      ii. Fit a multiple logistic regression model using only the training observations.

      iii. Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.

      iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

   (c) Repeat the process in Question (5b) three times, using three different splits of the observations into a training set and a validation set. In this question, keep the 70/30 training/validation set split. Please set `random_states` (Python) or `seeds` (R) of 2, 6, and 9 to obtain the three different splits of the observations. Comment on the results obtained.

[4] 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.

(d) Now consider a logistic regression model that predicts the probability of default using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Please use a 70/30 training/validation set split, and set `random_state=42` (Python) or set `set.seed(42)` (R) when splitting the data. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.