# Machine Learning - Problem Set 4

PPHA 30546 - Professor Clapp
Winter 2024

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Wednesday, February 28th.** 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 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 PPHA 30537/8 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 6 of the *Introduction to Statistical Learning* textbook:

   (a) Question 9, parts (a), (b), (e)-(g)

   - For (a), please use a 50/50 training/test split. To avoid confusion among partners and facilitate grading, please also set `random_state=37` when you split the data. Also, be sure to standardize the data (before performing PCR and PLS) using `scikit-learn`'s `StandardScaler` command. You should scale *after* splitting your data.

[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).

- For (e) and (f), be sure to use 10-fold cross-validation (10FCV) on the training set, shuffle the data randomly for splitting, and set `random_state=1`. Report the test error obtained, along with the value of $M$ selected by cross-validation, both by minimizing the appropriate cross-validated error and using the "elbow method."
- Python does not have a PCR command, so you should use `scikit-learn`'s `PCA` command, then run an OLS regression using the resulting principal components.
- `Scikit-learn` does have a PLS regression command (`PLSRegression`).

2. Do the following questions from Chapter 8 of the *Introduction to Statistical Learning* textbook:

   (a) Question 4

   (b) This question is a modified version of Question 9. It involves the `OJ` data set which is available on Canvas.[2]

      i. Create a training set and a test set. Please use a 70/30 training/test split and set `random_state=3` when you split the data.[3]

      ii. Fit a full, unpruned tree to the training data, with `Purchase` as the response and the other variables as predictors. Set `random_state=2` when calling the `DecisionTreeClassifier()` function. What is the training error rate?

      iii. Create a plot of the full, unpruned tree from the previous question.[4] The plot is a mess, isn't it? For the purposes of this question, fit another tree with the `max_depth` parameter set to 3 in order to get an interpretable plot. How many terminal nodes does the tree have? Interpret the information displayed in the first of the terminal nodes on your plot.

      iv. Use your fit of the full, unpruned tree to predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?

      v. Use cost complexity pruning to determine the optimal subtree for prediction by tuning the the $\alpha$ hyperparameter.[5]Produce a plot with the values of $\alpha$ (`ccp_alpha`) on the x-axis and the cross-validated classification error rate on the y-axis calculated using 5-fold cross-validation (5FCV).[6] Which $\alpha$ corresponds to the lowest cross-validated classification error rate?

      vi. Now produce a second plot showing the tree size on the x-axis and the cross-validated classification error rate (that you calculated in the method in the previous question) on the y-axis.[7] Which tree size corresponds to the lowest cross-validated

[2]For variable definitions, see https://rdrr.io/cran/ISLR/man/OJ.html.

[3]Note that there are some redundant predictors in the dataset. You can ignore this complication and use the full dataset for prediction.

[4]There are two ways to plot a tree in Python: `scikit-learn`'s `plot_tree()` function and `Graphviz`. The latter is a little difficult to work with, so use the former. The `scikit-learn` `export_tree()` function can also be useful if you want to produce a text summary of the fitted tree, but you're not required to do so for this assignment.

[5]This is the `ccp_alpha` argument in `scikit-learn`'s `DecisionTreeClassifier()`.

[6]Use the training dataset for 5FCV, shuffle the data randomly for splitting, and set `random_state=13`.

[7]Note that tree size is the number of terminal nodes or leaves and you can find this using the `.get_n_leaves()` method after fitting the model.

classification error rate?Briefly explain why the value of $\alpha$ affects the tree size and the classification error rate.

vii. Produce a plot of the optimal pruned subtree obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.

viii. Compare the training error rates between the pruned and unpruned trees. Which is higher? Briefly explain.

ix. Compare the test error rates between the pruned and unpruned trees. Which is higher? Briefly explain.