Machine Learning - Mini-Project 3

PPHA 30545 - Professor Clapp
Winter 2023

This assignment must be handed in via Gradescope on Canvas by **11:45pm Central Time on Wednesday, February 22nd.** 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 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 Overview

It is early in July of 2020, and you have just been hired to work for the Centers for Disease Control and Prevention (CDC).[2] As you probably know, the CDC is a national health protection agency

[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).

[2] Your family is very proud and all of your friends are jealous of your great gig. You tell them you're so glad that you're going to take Machine Learning at Harris in the future, as it somehow really helped you land the job.

tasked with protecting public health and safety by preventing and mitigating disease, injury, and disability in both the United States (US) and abroad. According to their website the CDC "conducts critical science and provides health information that protects our nation against expensive and dangerous health threats."

The number of COVID-19 deaths in the US has just surpassed 100,000. Things are looking grim around the world, but hope is on the way in the form of several potential vaccines. Your team at the CDC has been tasked with optimizing the distribution of a future vaccine to the places in the US that are most in need. To aid in doing so, you need a good prediction of which communities will be the most hard-hit when the vaccine is ready.

You colleagues have asked you to develop a model that predicts COVID-19 deaths per capita using socio-economic, health, and weather/pollution data from any and all available sources. Predictions based on your analysis will help shape your office's recommendations about how to best deploy limited vaccines. The project has two parts: understanding the available data and performing data analysis and answering questions.

2 Understanding the Data

The full dataset is available for download from Canvas, but you should familiarize yourself with how it was constructed. The data was compiled from the following sources:

1. NY Times: Daily data on COVID-19 cases and deaths at the county level comes from the New York Times GitHub Repository. Cumulative cases and deaths were calculated for counties in the continental United States as of July 1, 2020.[3]

2. Opportunity Insights: County characteristics, including baseline health measures, come from Bergeron, Chetty, Cutler, Scuderi, Stepner, and Turner (2016). The data are available for download on the Opportunity Insights Data Library webpage.[4]

3. PM COVID: Annual, county-level data on air pollution (PM2.5) and weather (average winter/summer temperature and relative humidity), and baseline mortality come from Wu, Nethery, Sabath, Braun, and Dominici (2020). The pollution and weather measures used in this project are calculated by averaging across annual averages from 2000-2016.[5]

[3] See https://github.com/nytimes/covid-19-data for additional details. Note that three jurisdictions (New York, NY; Kansas City, MO; and Joplin, MO) report their counts at the city level instead of by county. In each of those instances, the cases/deaths were divided evenly among the overlapping counties. Note also that the CDC has similar data that can be disaggregated by demographic characteristics (gender, age group, race/ethnicity), but access to that data is restricted and requires securing a data-use agreement.

[4] See the description of the "County Characteristics (Described in eTable 9)" dataset for additional details.

[5] See https://github.com/wxwx1993/PM_COVID for additional details.

3 Data Analysis[6]

1. The "Variable Description.xlsx" spreadsheet contains a list of variables that we'll use for our analyses. Note that this is not a full list of all the variables in the dataset, although it's close (we ignoring a few perfectly co-linear predictors). Filter the full set of variables in the dataset down to the the Opportunity Insights and PM COVID variables listed in the spreadsheet along with *county*, *state*, and *deathspc*.[7]

2. Compute descriptive (summary) statistics for the subset of Opportunity Insights and PM COVID variables you filtered in previous question.

3. Note that some variables have missing values. This causes problems when estimating the models. Normally we'd impute missing values by replacing them with their mean or median value, but to keep things simple, given the size of our data, you should drop all observations (rows) with missing values.

4. Create a separate dummy variable for each of the 48 states and the District of Columbia in the dataset (so you'll create 49 dummy variables in total, but dropping observations with missing values may reduce this number).

5. Split the sample into a training set (80%) and a test set (20%). Be sure to set a random seed so you can replicate your work.[8]

6. Using the training set, fit a model of COVID-19 deaths per capita ($y = deathspc$) as a function of the Opportunity Insights and PM COVID predictors listed in the spreadsheet, as well as state-level fixed effects (the state dummy variables) using OLS.

   (a) Using the model you fit, calculate and report the *MSE* in both the training and test sets.

   (b) Why might you be concerned about overfitting in this context? Is there any evidence of overfitting? Briefly explain.

7. Use the training set to estimate ridge regression and the lasso analogs to the OLS model in the previous question. For each, you should report a plot of the cross-validation estimates of the test error as a function of the value of the hyperparameter (λ) that indicates the tuned value of λ. Hint: to do so you should be sure standardize your predictors and tune the hyperparameter by:

[6] Note that your responses in this section should be submitted and will be graded, but that some of the "questions" that follow explain how you should prepare the data and don't require that you submit anything.

[7] Hint: rather than typing each variable name individually to select the variables to be summarized and for modeling, you can simply import them from the Excel file, add them to a list, and subset the data from that list. This will be covered in lab. Also, note that the *county* variable is just an index to give real-world context to each observation. It will not be used for modeling.

[8] There are odd patterns in the results with some splits of the data, so to avoid confusion and facilitate grading, please set `random_state=25`.

(a) Calculating each model for a grid or range of values of λ. You'll want to adjust the values you use based on the data, but start by using 100 values of λ from 0.01 to 100.[9]

(b) Using 10-fold cross-validation (10FCV) (on the training data) to estimate the test error for each model at the given value of λ.

(c) Plotting the 10FCV estimates of the test error as a function of the value of λ.

(d) Choosing the optimal value of λ.

(e) Re-estimating your model using that optimal value of λ.

8. Using the ridge regression and the lasso models you trained based on the optimal values of λ you found in the previous question, calculate and report the training and test set prediction errors (*MSE*) for each model. Did ridge regression and/or the lasso improve your prediction over OLS? Which model performs the best? Briefly explain which model you would recommend to the CDC and why.

[9] Hint: to do so, start by creating a range of 100 numbers from -2 to 2. You can do that by starting at -2 and incriminating each subsequent number by $\frac{1}{25}$. Let a denote those values and b denote the values of the grid for tuning λ. You can calculate $b = 10^a$.