Machine Learning – Lab 3

PPHA 30545 - Professor Pouliot

Winter 2022

This assignment must be handed in via Gradescope on Canvas by 11:59pm Central Time on Monday, February 28th. You are welcome (and encouraged!) to form study groups (of no more than 3 students) to work on the problem sets and labs 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.

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. For the latter, you may type your answers or write them out by hand and scan them (as long as they are legible).

You are allowed to consult the textbook authors' websites, 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 under the guise of tutoring.

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).[1] The CDC is a national health protection agency 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 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.

[1] 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 next year, as it somehow really helped you land the job.

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.[2]
2. Opportunity Insights: County characteristics, including baseline health measures, come from Bergeron, Chetty, Cutler, Scuderi, Stepner, and Turner (2016).[3] The data are available for download on the Opportunity Insights Data Library[4] webpage.[5]
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).[6] The pollution and weather measures used in this project are calculated by averaging across annual averages from 2000-2016.[7]

3 – Data Analysis

1. The "VariableDescription.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 Opportunity Insights and PM COVID variables listed in the spreadsheet along with 'county', 'state' and 'deathspc'.[8]
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.

[2] 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.

[3] https://jamanetwork.com/journals/jama/article-abstract/2513561

[4] https://opportunityinsights.org/data/

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

[6] https://www.science.org/doi/10.1126/sciadv.abd4049

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

[8] 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.

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).
5. Split the sample into training (80% of the data) and test (20% of the data) sets. Be sure to set a seed so you can replicate your work.
6. Using the training data, estimate the relationship between COVID-19 deaths per capita (*y = deathspc*) and 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. Based on those estimates, calculate and report the *MSE* and $R2$ in both the training and test sets.
    b. 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:
    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 set) to estimate the test error for each model at the given value of λ.
    c. Plotting the cross-validation 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 optimal values of λ you found for Ridge Regression and the Lasso in the previous question, calculate and report the training- and test-set prediction errors (*MSE* & $R2$) for each model. Did Ridge Regression and/or the Lasso mitigate overfitting? Briefly explain your results.
9. Bonus question: now compare the test-set prediction errors from Ridge Regression and the Lasso to that from OLS. Is this what you expected? Briefly explain.

[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 1/25. Let a denote those values and b denote the values of the grid for tuning λ. You can calculate b= 10^a.