Statistics 215A Final Project Fall 2020

Due: Friday December 11 at 11:59 PM

Note, there is no HW with this lab. Please push a **lab_final/** folder with the following files by the deadline:

- **lab_final.Rmd** or lab_final.Rnw: the raw report + code.
- **lab_final.pdf**: the output of lab_final.Rnw/lab_final.Rmd. This output should be no more than 12 pages and should not contain any code.
- **R/**: a folder containing any .R scripts (e.g. load.R and clean.R) that will be sourced in lab_final.Rmd and any other pieces of code you use.

You may create additional folders if you wish, but please keep your project directory organized. Push everything necessary to reproduce your report, but nothing else (e.g., do not push your **data/** or **documents/** folders). Note also that you do not have to create blinded files.

1 Introduction

The COVID-19 pandemic has been an ever-present part of our lives in the US and across the world for most of this year. It has caused tens of millions of infections, over 1.3 million deaths, and in the US we currently find ourselves in a dangerous third wave of the pandemic. In March, the Yu Group responded to a call for data science expertise by the one-week old non-profit Response4Life (R4L). They needed help to distribute personal protective equipment (PPE) by identifying hospitals that were in greatest need of PPE across the US (please read "Getting the right equipment to the right people" for more info). This partnership required many ER data science skills including getting data, calling hospitals, using personal social networks to understand PPE supply logistics chain, and organizing volunteers. This partnership also drove Yu Group's statistical machine learning work. They decided to forecast demand at county level (since no hospital data was available) for PPE with enough lead time for the R4L logistics team to deliver PPE supplies when and where they were needed most. For example, our partners at R4L told the Yu Group that their distribution chain required at least 5 to 7 days for deliveries, and this information informed the horizon of our short-term forecasts.

While the PPE situation was particularly dire in the beginning of the pandemic, we find ourselves now in yet another difficult phase marked by supply scarcity[1]. This is the domain problem that this lab is aimed at solving, using the data curated by the Yu Group on GitHub or additional data that you can find and document. Your task is to help the California state government prioritize the distribution of personal protective supplies or PPE to those counties that are most in need, within the same supply chain constraints that the Yu Group faced when working with R4L. Your group will focus on two geographic areas in order to inform the PPE needs of diverse types of counties by providing a better understanding of the course of the pandemic in urban, exurban, and rural areas in both the north and south of the state.

The first cluster is the following 9 Bay Area counties: **Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma**.

[1] https://www.nytimes.com/2020/11/17/magazine/n95-masks-market-covid.html

In addition to the Bay Area, you should choose one of the two following clusters of counties to consider:

- Counties bordering Los Angeles: **Kern, Orange, San Bernardino, and Ventura**.
- Southern Central Valley counties: **Fresno, Madera, Merced, Tulare, and Stanislaus**.

2 Data

In the stat-215a-fall-2020 repository, as of the release date of this lab you have the time series of case and death counts for every US county up to November 19, as well as demographic and health-care features. Every day, the repository will be updated with the previous day's recorded counts, which you can use to test your forecasts in real time.

The data for this lab comes from the Yu Group's covid19-severity-prediction repo. There are three datasets, all of which contain information at the county level:

- counts.csv: Cumulative case and death counts time series from USAFacts.
- county_data_abridged.csv: Demographic, socioeconomic, and health-care data, social distancing and mask-use statistics, and geographical identifiers. The columns are described in the file list_of_columns.md found in the covid19-severity-prediction repo.
- county_adjacency2010.csv: Neighboring county pairs.

You may use additional data if you wish, but please discuss its provenance and how it can be accessed (e.g., you could include a link to a website or a Google Drive link, but do not upload the data to your repo).

3 Tasks

With the goal in mind of helping the CA public health department distribute PPE, you will work with both the case and death counts as your outcomes. You should try to predict both outcomes, and connect your forecasts back to the domain question. The exact from of the two outcomes are up to you. For example, do you want to work with cumulative counts from the beginning of the pandemic? Starting from a later day? What about new counts, or an average over some sliding window? Will you regularize by population sizes? State clearly what pre-processing steps you apply to the outcomes, make a clear argument as to why you make the choices you do, and explain how your choices relate to the domain problem and corresponding statistical problem.

To evaluate your forecasts, use all three metrics defined at the beginning of Section 5.1 of [Altieri et al. 2020] (included in the lab_final folder in the stat-215a-fall-2020 repo). You may also consider additional error metrics that you have good reasons for (document the reasons), but your GSI will use these three metrics to evaluate your predictions for the future data.

This project involves the following tasks:

1. Produce a collaboration plan. How will your group divide up the tasks in this report? Who will do what? How often and when will you meet? You should discuss this and write it up during your first group meeting. The write-up of your initial collaboration shouldn't change plan even if you decide to do things differently as you progress.
2. As a baseline, implement the “linear" predictors and one of the exponential predictors described in Section 3 of [Altieri et al. 2020] to predict cases and deaths.

3. In [Altieri et al. 2020] the authors use a method they call CLEP (Combined Linear and Exponential Predictor) to combine their five predictors into a single ensemble predictor based on recent performance. CLEP's weights are based on a method originally developed in [Schuller et al. 2002]. Implement CLEP for the two predictors from the previous task.

4. Now implement at least two predictors of your own that are distinct from those used in [Altieri et al. 2020]. If you wish, you may use different pairs of predictors for the different clusters of counties (Bay Area and the Central Valley or Los Angeles area counties), but you should use the same predictors for all counties within a cluster. Discuss your data-splitting scheme and how you tune any hyperparameters your methods may have.

5. In [Altieri et al. 2020], the authors do not tune the parameters of CLEP, but instead use $c = 1$ and $\mu = 0.5$ (equations 3.7 and 3.8 starting on page 18). Your next task is to use a similar weighting scheme with your predictors from the previous task (and the linear and exponential predictors if you wish), but tune the parameters c and μ and do so separately for the Bay Area counties and your second set of counties. Include $c = 1$ and $\mu = 0.5$ as well as at least two additional values each for both c and μ (resulting in nine hyperparameter combinations). Again, discuss your data-splitting scheme and how you tune the hyperparameters.

6. Evaluate your methods and the ensemble method from the previous step, along with CLEP and the individual linear and exponential predictors from [Altieri et al. 2020]. Which methods perform the best? What can be learned about the methods from the ensembles? Are there differences in the two clusters of counties? What about within each geographic area?

7. Evaluate your best model further. Check the fit of the model over a recent one-week period. Are there any outliers? How much do the fitted parameters vary from one day to the next? What about across the counties and geographic regions? What are the important features in the model? Discuss the stability of your prediction results and the models.

8. For the same one-week period, produce maximum absolute error prediction intervals (MEPI) based on your best method's past performance as in section 4.1 of [Altieri et al. 2020]. Alternatively, you may use a different method of your choosing to produce the prediction intervals. For what percentage of the days do your intervals cover the true value?

9. For every Bay Area county as well as those in your second set of counties, create point predictions and intervals for December 10 to December 16 and save them to a file called `5-day-ahead.csv`. This file should be in the same format as the `5-day-ahead-example.csv` file that was included in the `stat-215a-fall-2020` repo. The last day of data that will be added to the repository is Wednesday, December 9, which should be available on the morning of Thursday, December 10.

4 Discussion

Address the following:

- Discuss your predictions of cases vs. the predictions of deaths. Was predicting one more challenging than the other? Do you trust your methods and predictions? Why or why not?

- Do you think your methods will work for other times during the pandemic? In the past? What about the future? Explain.

- Discuss your initial collaboration plan. Did you end up following that plan or did things change as the project progressed? What worked and what did not? What parts of this collaboration did you find most challenging? Is there anything you would have done differently?

5 Notes

1. Think carefully about training and validation. The test set is December 10 to 16, but you will also probably want to get an estimate of future error using the data you have available to you. The new daily data could be useful in this regard.
2. There might be many parameters to tweak among the methods and weighted ensembles, so you should consider using the SCF clusters and parallelization. If you do, please be sure to include all code necessary to reproduce your results including SLURM scripts.
3. We purposely left this lab more open-ended, so feel free to get creative, but also try to be realistic about what you can accomplish in three weeks. If you run into road blocks please don't feel the need to start over or discard your hard work. Document what you've done, discuss the obstacles you've encountered, and explain what you could have done differently to avoid the trouble you're having.
4. As this is your final project, please note that I will allot points towards visualizations, readability of code, thoughtfulness of your introduction/discussion/conclusion, overall formatting and presentation (e.g., cross-referencing of figures and tables, appropriate captions, no R output), and reproducibility in addition to the guided analysis above.

6 Bonus opportunity

For up to 5 bonus points for each group member, follow the instructions below to call a California county office and try to verify the data we have for that county. If you want to participate, please let your GSI know via Slack or email and they will assign a county to your group. Write a short description of the conversation in an addendum to the report.

You may use the following script:

Hi, I'm a student at UC Berkeley. I'm calling to ask about your county's COVID-19 data reporting practices. Specifically:

- Where does the data come from?
- How do you determine whether or not to attribute a death to COVID-19?
- How do you decide what day to report the death?
- Can you share the data with me?

If you are able to get data from the county, you may also compare that data with the data from USAFacts that you used in this lab and describe any differences that you find.

References

[Alt+20] Nick Altieri et al. "Curating a COVID-19 Data Repository and Forecasting County-Level Death-Counts in the United States". In: *Harvard Data Science Review* (Nov. 3, 2020). DOI: 10.1162/99608f92.1d4e0dae. URL: https://hdsr.mitpress.mit.edu/pub/p6isyf0g.

[Sch+02] Gerald Schuller et al. "Perceptual Audio Coding using Pre- and Poster- Filters and Lossless Compression". In: *IEEE Trans. Speech and Audio Processing* (6 2002), pp. 379–390.