EE 510: Machine Learning Theory & Algorithms **Spring 2020**

Mini Project 2

Due: May 15, 2020, 11:59PM PT

Student Name: *Instructor Name: John Lipor*

Introduction

COVID. Need I say more? While the IHME model [1] has been very popular among politicians and journalists, it has also received a fair bit of criticism for a variety of reasons. An alternative model, with less appealing visualizations, is that of [2], in which a lone data scientist utilizes machine learning to learn the parameters of an infectious disease model. The result is recognized by the CDC [3] and has shown to be among the most accurate existing approaches to forecasting. While reproducing this approach would be an excellent project (that I encourage you to try in your free time if you are bored), it would also require more time than is reasonable for this course.

For this mini project, your task is to submit an entry to the COVID19 Global Forecasting (Week 5) Kaggle challenge [4]. The forecasting problem can be viewed as a time series prediction problem. While we have not studied time series data, this problem can be easily cast as a supervised regression problem (see e.g., [5]). For this project, you are free to use any algorithm in `sklearn` *except* those based on neural networks[1] as well as `xgboost`. You are also permitted (by both myself and the competition) to use external data sources. For example, one could attempt to model the model [2] as part of their predictor (though I have no idea if this would be a good approach), or incorporate mobility or other data from the numerous COVID datasets on Kaggle.

Notes:

1. DSS rules apply for all mini projects.
2. Be sure to cite any sources you used, including/especially Kaggle notebooks.
3. I am unsure whether there will be a Week 6 challenge. If such a challenge does arise next week, feel free to submit to that one either additionally or instead of Week 5.

Requirements

You must create a report in either LaTeXor a Jupyter notebook that contains the following sections:

1. Problem description
    - Explain what data you have to work with, what algorithms you will use, what your goal is, and why anyone should care.
2. Exploratory data analysis (EDA)
    - There are many excellent tutorials on EDA online. Be sure to cite whichever ones you found helpful and comment on them in Sec. 6.

[1] There is nothing wrong with neural networks, but there are several other classes on campus devoted to their study. Further, tuning their parameters can be a time sink that will get in the way of your learning for this project.

- Be sure to make any significant findings stand out and try to keep this concise.
- **(NEW)** Since you are working with geospatial data, your project must include at least one visualization overlaid on a map, either of the world or of one specific country. This could be used to look for visual trends in the data (in the EDA section) or after prediction to determine whether any regions of the world are predicted to experience high infection or death rates in the near future (in the Evaluation section). Several existing Kaggle notebooks provide such visualizations.

3. Challenges
    - What were the challenges you encountered when applying machine learning to this dataset?
    - Did these challenges mainly result from the data? From results? Installing libraries to perform preprocessing?
4. Approach
    - Provide a thorough but concise description of your approach, including (but not limited to) your approach to wrangling, preprocessing, and improving the performance of your predictor.
    - This should be much clearer than what is typically found in Kaggle notebooks, but make sure it stays concise. A summary paragraph at the beginning of this section could be helpful.
    - **(NEW)** Since we have now formally studied model selection, you must use either a validation set or cross validation to tune your model parameters. Be sure to describe how you went about this and why.
    - **(NEW)** Be sure to clearly state whether you considered any external data sources (this is permitted in the competition).
5. Evaluation and summary
    - Make use of any meaningful metrics (e.g., classification error, precision, recall, confusion matrix, f1 score), but only include metrics that tell you something different.
    - Consider diving into the data to see if you can determine any trends regarding which points were misclassified.
    - Likewise consider if there were any features that were particularly important or unimportant.
    - Summarize your solution, describing what worked, what didn't, what the main limitations are, and any general conclusions about the dataset.
6. What I learned
    - Describe the main skills/tools that you learned and used for this project and how you learned them.

Grading

The grading breakdown is below. Each section will be graded according to technical correctness, effort, and creativity. Note that clarity of writing is a major component. You should put yourself in the place of writing for a boss or senior in a workplace. If your writing is terrible, they will assume your work is terrible.

Item	**Percentage**
Description	5%
EDA	20%
Challenges	10%
Approach	30%
Evaluation	20%
What I learned	5%
Clarity/conciseness of written communication	10%

References

[1] (2020) Covid-19 projections. [Online]. Available: https://covid19.healthdata.org/united-states-of-america

[2] (2020) Covid-19 projections using machine learning. [Online]. Available: https://covid19-projections.com/

[3] (2020) Covid-19 forecasts. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/forecasting-us.html

[4] (2020) Covid19 global forecasting (week 5). [Online]. Available: https://www.kaggle.com/c/covid19-global-forecasting-week-5/overview/description

[5] (2016) Time series forecasting as supervised learning. [Online]. Available: https://machinelearningmastery.com/time-series-forecasting-supervised-learning/