Early Detection of COVID-19 Hotspot Counties Using Data Science

Problem: Quarantine measures and early testing of COVID-19 are necessary for containing community spread of SARS-CoV-2, and such steps are most effective when taken at early stages of community outbreaks. Therefore, early and accurate detection of community outbreaks is critical to address the threat of resurgent waves of COVID-19.

Solution: The purpose of this project was to design **a detection tool** utilizing state-of-the art data science methods for advanced detection of potential community outbreaks that can reduce community spread and save lives.

- To this end, an early detection algorithm should be able to analyze data at a very granular level **in the presence of limited data** and rapidly pick up future trends.
- From the practical perspective, an algorithm that can accurately and quickly detect potential outbreaks early on is desired.

- On one hand, focusing on local county level analysis may result in severe overfitting due to relatively sparse data, especially during early stages. On the other hand, expanding the analysis to a less granular state level results in less noise but may be extremely biased if the counties were clustered poorly.

- We hence provide a solution that **robustly forecasts and clusters** sparse time series data, with **high throughput even on modest computing hardware**. Cross validation on historical data also demonstrated that our solution was able to detect incoming hotspots a week in advance

Methodology: This project develops **a novel two-step estimation scheme** to balance this:

1. We automate the feature engineering on the cumulative confirmed case numbers, and detrend the data. Using the past two weeks at every time step, we fit an exponential growth model and use time-delayed embedding to generate much richer features and less noisy estimations of the historical disease growth rate and outbreak time of each county
2. To reduce the estimation variance and make the problem tractable, we approximate the **mixed integer convex program (MICP) formulation** to classify these counties, based on the similarity of their model fits in the first stage, into different clusters and simultaneously re-estimate the disease growth rates and the outbreak times of these clusters with XGBoost. In particular, the number of clusters is determined by cross validation to balance the **bias-variance tradeoff**.

Impact: The final output is thus **a classification of counties based on their likelihood of future COVID-19 outbreaks** that can be obtained on a personal computer within a few minutes. In particular, back testing on historical data demonstrated that our solution was able to **detect incoming hotspots a week in advance.** As such, this method would provide timely warning to policy makers when a county starts to move from a “low risk” cluster to a “high risk” cluster so that they can intervene ASAP.

Release: The Python code can be found in our GitHub Repository at https://github.com/Runespear/COVID-tracking