CSE 8803 EPI: Data Science for Epidemiology, Fall 2022

Lecturer: B. Aditya Prakash
Scribe: Bhavay Aggarwal, Shubham Agarwal

Month 10, 2022
Lecture 15 : FORECASTING I

1 Summary

So far we have looked at methods of modelling diseases, detecting outbreaks and surveilling its spread. The next step here is to use our modelling of the network and surveillance techniques to try to predict the future of the disease spread. Predicting disease spread is essential for making informed public policy decisions and taking actions to minimize the impact of the disease on the population. While the influx of large data collection sources has accelerated research into real time epidemic forecasting, it is imperative to incorporate the different causal factors while forecasting to ensure accurate predictions.

Epidemic Forecasting can be split into different categories based on the tasks in hand, indicators of interest and the spatial/temporal scale. These categories also dictate how we evaluate our model's forecasts. We further look into hybrid models which combine the previously discussed mechanistic models with machine learning methods. These models are able to efficiently utilize the variety of datasets available while incorporating the domain-based prior knowledge of mechanistic models.

2 Why Forecasting?

Regularly looking at the global cases graph and trying to predict when the pandemic will die out, the covid-19 pandemic has brought out the armchair epidemiologist in us all. Forecasting diseases is very similar to weather forecasting in the sense that it gives the people a sense of what to expect. Additionally, it allows governments to prepare medical resources and make informed public policy decisions. The importance of forecasting makes it crucial that it is accurate and incorporates the different causal factors.

The causal factors mainly are -

- Current Number of Infections
- Interventions in place
- Contract patterns
- Exposure to disease

The advent of social media and communication technologies has brought along vast amounts of data which at first sight might seem unrelated to epidemics but in fact are very useful. Google Search Trends, Facebook's Covid Impact Survey and Safegraph's mobility data are just some of the examples of different types of data available and are an *indirect* measure of the causal factors. Advancements in machine learning have also made models capable of ingesting such data readily available. Example citation [1].

3 Epidemic Forecasting

3.1 Forecasting Tasks

Forecasting tasks can be broadly seperated into three categories -

- **Real Valued Predictions** - look at epidemic indicators like mortality rate and cases to try and understand the trend of the disease. Delays in reporting data make nowcasting useful. Short term forecasting helps increase preparedness and Long term forecasting although hard, can help us understand how the epidemic is going to pan out.
- **Event based Predictions** - look into events such as peak-time and onset time which are indicators of the intensity of the epidemic. These events are used as signals for interventions like shutting of schools and vaccination drives.
- **Epidemiological indicator Predictions** - look into the composite indicators that characterize the behavior of the epidemic like the reproduction number and the final infected size.

3.2 Targets of Interest

The epidemic growth can be analysed by looking at indicators such as cases mortality and hospitalizations. These indicators although might not always be accurate because during the covid season, Influenza %ILI might get mixed with symptomatic covid outpatients so additional indicators like lab-tested hospitalizations would be needed.

3.3 Spatial and Temporal Scales

Forecasting of epidemics is done on different spatial granularities. Regions are grouped together to make monitoring and reporting more easier. Additionally, forecasts are made for different times scales like weekly and daily.

3.4 Model Evaluation

The next most important step after training models is to evaluate them. We want to measure the success of our predictions and the success might be different in different forecasting scenarios. Forecasting outcomes can be split into two categories -

- Point Forecasts
- Probabilistic Forecasts

Point Forecasts have single valued results like number of infections or deaths. Metrics used for point forecasting include RMSE, MAE and MAPE. RMSE and MAE measure the error in L2 and L1 norm respectively whereas MAPE allows us to get the ballpark of our prediction using the prediction error. Since, these forecasts are of great social and economic significance, it becomes essential that we are certain about our predictions. Point forecasts however are not able to capture the confidence of the model and this is why probabilistic forecasts are preferred. Probabilistic forecasts capture the uncertainty of model predictions by using confidence intervals while also considering the accuracy of the model. Log Score

calculates the binned log probability of the ground truth, Interval scores are used to penalize how far the models predictions are from the ground truth. Additionally, we want to penalize flat distributions because that means the model is not confident in predicting and assigns equal probability to all outcomes. We also want to penalize if the ground truth is lower than the lower bound and greater than the upper bound. Coverage score is another metric which measures the fraction of times the ground truth actually lies in the confidence interval or in simple words penalizes an overconfident model. More recently, researchers have adopted the Weighted Interval Score(WIS), which aggregates interval scores for multiple intervals and aggregates them. WIS also builds upon log score in the sense that it is unbounded, and especially for forecasting diseases like covid-19, it is a more suitable metric.

4 Modeling Paradigms

[2] talks about different type of paradigms, dataset used by them, the type of task and features used.

4.1 Mechanistic Models

These models have been thoroughly discussed in the previous lectures. Mechanistic models are the workhorses in epidemiology. The population is divided into different compartments based on the based on the disease state and they move between compartments based on the disease progression. ODE's, Metapopulation models and Agent based models are the primary examples. These models require domain knowledge and its parameters require intensive testing for sensitivity.

4.2 Statistical/ML Models

Advancements in machine learning methods has lead to a large scale collection and maintenance of publicly available datasets. Newer architectures have been successful in finding patterns even in complex data forms. Due to their success a large number of optimization algorithms are also available. Some approaches are regression based, language and/or vision models, neural networks and density estimation models. ML Models are useful as they can handle different types of datasets like Languages, Images, time-series etc.
The idea here is to find a function that can forecast the target based on the input data, however, do note that this would generate an approximate forecast. This is achieved by minimizing the comparison between the prediction from the function and the ground truth. The comparison is called Loss and the function that generates the comparison is known as a loss function. And the goal is to minimize the loss generated by the loss function.

4.3 Hybrid Models

Given the wide spread use of ML models, a ML engineer doesn't need to have domain knowledge to apply a model on epidemic data set - this might work is certain scenarios but in most scenarios domain knowledge combined with ML model would produce better and reliable results. Hybrid Models combine advantages of both Mechanistic Models and ML Models. They use domain-based priors and expert knowledge from mechanistic models and

flexible data-driven approach from statistical/ML methods. A model could use statistics to estimate mechanistic parameters or wisdom of crowd with ensemble models.

References

[1] B. Adhikari, X. Xu, N. Ramakrishnan, and B. A. Prakash. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 577–586, 2019.

[2] A. Rodríguez, H. Kamarthi, P. Agarwal, J. Ho, M. Patel, S. Sapre, and B. A. Prakash. Data-centric epidemic forecasting: A survey, 2022.