Probabilistic Models for Predicting COVID-19 Pandemic Spread

Akash PS * Kethan M V * Venkatesh K * Suhas K Mentor - Prof. Nitin V Pujari

Department of Computer Science and Engineering, PES University.

Problem Statement

Developing and Deploying probabilistic models to preditct the COVID-19 pandemic spread.

The World Health Organization (WHO) announced that coronavirus disease (COVID-19) has spread throughout the world and has been declared as a pandemic on March 11th 2020. The world has placed epidemic modelling at the forefront to enable the world and its stakeholders to take informed decision making in containing the spread.

Background

There is a strong correlation between population density and the COVID-19 cases confirmed
Exponential Smoothing is known to be the best model for forecasting.
Piecewise Linear regression has relatively less error and the data points are close close the regression line.

Dataset and Features / Project Requirements / Product Features

Data is taken from open source website "covid19india.org" with file "Districts.csv". Data consists of 238239 records with 8 features.

The data with features "Date","District" and "Confirmed cases" is considered for model training and testing.

User Interface(UI) is created to visualize the pandemic spread by inputting the district of interest and model of user's choice.

Design Approach / Methods

Of all the districts data available, five districts are chosen for training and testing the model.

Four probabilistic models are developed and deployed to predict the number of confirmed cases for a particular distict on a particular date.

Four different training sets is created for each model and the best model among them for each district is chosen for prediction.

Accuracy of individual probabilistic models are recorded and best performing model is reported.

Results and Discussion

Piecewise Linear Regression Model has the highest average accuracy among all the four probabilistic models

All models except hierarchical bayesian model reported more than 90% Accuracy for few districts.

Accuracy of Simple Exponential model decreases drastically over time.

Hierarchical Bayesian model follows a lognormal curve so the model wouldn't be able to predict the results accurately as we have seen in the case of Delhi and Mumbai due to the steep rise in the confirmed cases.

Hierarchical Bayesian models tend to underestimate the number of confirmed cases which is true as the model fitted assumes that the cases follow a lognormal curve.

XGBoost provides high accuracy for the first 3 days of prediction. The accuracy reduces as we move towards future days.

The drawback of this model is that it does not predict according to the trend hence will not give satisfactory results for dates more than a week. For instance the percentage of incorrect prediction for confirmed cases is almost 20% for test data after 7 days.

Summary of Project Outcome

Identifying appropriate research papers for the topic of interest.

Gathering informaton from open source websites and leveraging tools.

Predicting more than one value for simple exponential smoothing.

Converting regression models into probabilistic models by introducing confidence interval.

Learnt to train and test forecasting and probabilistic models.

Conclusions and Future Work

Finding the optimal number of segments for piecewise linear regression.

Reducing the confidence interval window for Piecewise Linear Regression model and Simple Exponential Smoothing for better precision.

Checking the accuracy for Hierarchical Bayesian model using different curves like exponential curve.

References

[1] Hemant Bherwani, et al "Understanding COVID-19 transmission through Bayesian probabilistic modeling and GIS-based Voronoi approach:a policy perspective", 1 July 2020
[2] Andrea L. Bertozzi, et al "The challenges of modeling and forecasting the spread of COVID-19," July 2020.


Akash P S


Kethan M V


Suhas K


Venkatesh


Prof. Nitin V Pujari