

1 Introduction

2 Modeling

2.1 Data Preprocessing

Data is taken from the Covid Tracking Project [1], which is updated once per day. First, we add a column which takes the date formatted as YYYYMMDD and converts it to `datetime` format. Then we, find the minimum date in the dataset and find the difference in days between each row and that minimum date. That becomes our "days since start" category.

Covid Tracking Project assigns grade labels to the data, with A+ being high quality data, B being lower quality, etc. We prune all data which is not 'A+'.

We look at these features to model subsequent deaths: days since start, positive case increase, death increase, currently hospitalized, hospitalized increase, currently on ventilator, and in ICU currently. We remove all rows which do not contain all of these predictive features, and keep only these columns. Finally, we list each state and its number of row records after pruning.

A model is created for each state, because it's unlikely that e.g. the number of hospitalizations in New York would be useful in predicting the number of subsequent deaths in Colorado. Because some states do not provide detailed accounting of predictive features, we allow the user to choose which states to model, bearing in mind that states which don't provide detailed predictive feature data will result in a less accurate model.

Thus the data which is provided to the model is:

Feature	Type
Days Since Start	int
Positive Increase	int
Death Increase	int
Hospitalized Currently	int
Hospitalized Increase	int
On Ventilator Currently	int
In ICU Currently	int

Table 1: Prediction Features

2.2 Inputs and Outputs

Each model input x is a tuple of the features shown in table 1. Each model output y is a probability density of functions which map these features to a death increase after p subsequent days.

2.3 Kernel

2.4 Assumptions

3 Results

References

- [1] *The COVID Tracking Project*. The COVID Tracking Project. URL: <https://covidtracking.com/> (visited on 09/06/2020).