Decision Tree Report

Manasvi Sagarkar- 20CS60R28
Shreyas More - 20CS60R34

October 2020

1 Introduction

We worked with the COVID-19 dataset. We were given global data which included the date when this data point was collected, country name, number of confirmed cases, number of recovered cases and number of deaths due to COVID-19. We had to predict the number of deaths due to COVID-19.

2 Tree Construction

We start with considering the entire dataset at our root node. We initialise the variance of the node to infinity and use the minimum variance to decide which attribute to split the data on. We consider each attribute in turn and for all the data points, consider the split of the data that minimises the variance of the node. The attribute which can produce the best split (minimises variance of original node the most), is chosen as the attribute to split on. We can keep splitting the dataset in this manner until we reach a leaf node. We recognise leaf nodes as those for whom the variance cannot be reduced further by splitting. We found our tree implementation to run too slowly to use for our further experiments so all our experiments are on a scikit-learn regression tree.

3 Feature Selection

We had two attributes in our dataset that needed modification- one was the date attribute and the other was the country name attribute. For date, we extracted the month and the day from the date. All the years in the data were 2020, so we didn't use the year as a feature. For the country names, we considered two schemes to convert them into numerical features. One was to map each country to a numerical code and the other was to use a one hot encoding on the country names. Using one-hot encodings expanded the number of attributes we had since each country's encoding was now a separate attribute. We looked at two measures to decide which features to use. We considered 10 random 80-20 splits of data and compared accuracy and error for each choice of features subset based on the best values of accuracy and error for each subset. We used the R2 score as a measure of accuracy with 1 being the R2 score for a perfect decision. We also used the Root Mean Squared error of the predicted number of deaths. Based on these two values, we considered various choices of features and realised that dropping the ”Days” feature (which day of the month) which we extracted from the date lead to a slightly improved r2 score and reduced the RMS error across 10 random splits. Using one hot encoding for country names also performed better (on both r2 score and rms error) than using country codes. So we eventually used month, one-hot encodings of all countries, number of confirmed cases and number of recovered cases as our input features to predict the number of deaths. With these set of features, before pruning, our tree has the the best r2 score of 0.9997 and RMS error of 157 on 10 random splits of the data.

4 Depth of Tree

We plotted depth vs r2 score and depth vs RMS error graphs for depths from 1 to 120. We find that after a depth of around 20, our accuracy attains its best value and remains constant with slight fluctuations until depth of 120. Similarly, our RMS error reduces drastically after a depth of around 20 and has only relatively small fluctuations after this depth until the depth of 120. Having a very deep tree only increases chances of our model overfitting so a depth of around 25-30 would be a good depth to use for this model.

5 Pruning the tree

We employ a post pruning to improve the accuracy of our regression tree and it predicts the total deaths more accurately. Thus, we make use of cost complexity pruning technique to find the weakest link in the tree and prune it. The first step in cost complexity pruning involves calculating the sum of squared residual for each tree. Alpha is the tuning parameter such that nodes with the smallest effective alpha are pruned first. This value of alpha is calculated by performing k fold cross validation. Also, as alpha increases. more of the tree is pruned. Finally, the alpha value with minimum average error is chosen and the corresponding tree is the final regression tree.