MSAN 621 - Homework 3

Andre Guimaraes Duarte

December 1, 2016

For this assignment, I implemented a random forest machine learning classification algorithm to predict average life expectancy in a given country for a given year.

This document will explain the process through which I arrived at the final model included in `a3.py`. Note: this file only contains the final model that I used, and none of the intermediate steps (explained in the document below) are shown within the code.

What algorithms or approaches did you use for your implementation? Why did you choose the one you did? What hyperparameter(s) did you choose, and why?

I started by taking only the country and the year as features, with the intent of adding extra variables to try to improve prediction error. The life expectancy file provided for this assignment contains, at this point, all the necessary information to build a prediction model.

First, missing values within the year-range of the file are inputed linearly. Data for the extra years (1950-1959 and 2010-2016) are completed through linear extrapolation.

Then, country names are converted from categorical to numerical through a label encoder. This means that the country name can now be used as a feature for training.

The table is then melted: instead of having one row per country, I now have a column with countries, one with years, and the third has the life expectancy for every combination of (country, year) in the data. This gives me the X and Y matrices that I can directly use for model building.

Several models were tested, and their performance was assessed through 10-fold cross-validation (equivalent to an 90-10 split).

First, I tried using a simple linear regression model. The resulting MSE was 117.20, which is obviously too high.

I then tried Lasso and Ridge regression, but since only two features are in the model, neither model improves much over regular linear regression, and their MSEs are similar to the previous value shown above. I expect these methods to perform slightly better once more features are added to the system.

Then, I used a decision tree regression with an optimal depth of 30 (determined by lowest cross-validation MSE for several intermediate depths (from 5 to 50)). This gave me an MSE of 1.43. This is a considerably better result than linear regression.

After this, I used a random forest regression model, with 50 estimators (determined by lowest cross-validation MSE for several intermediate estimators (from 5 to 400)). This gave me an MSE of 0.96.

Therefore, I concluded that a random forest regression model was the most adequate for this exercise.

If you added data for features, what are the sources of the data and why did you add the data from these sources?

Having chosen a prediction model, I then proceeded to introduce extra features in order to get even better predictions. In general, I tried to include features that would intuitively correlate with life expectancy.

First, I incorporated the GDP data also provided to us for this assignment. Missing GDP values were set to 0. However, the resulting MSE for the same random forest model was higher than before at 11.04.

From the world data bank website, I also got data concerning HIV rates, income per capita, and internet usage in all countries. Missing values for HIV rates and internet use were backfilled with zeros, and forward filled by linear extrapolation (indeed, neither existed before 1960). For income data, missing data were filled through linear interpolation. Adding these extra features to the model increased the MSE even more to 17.64.

Therefore, I concluded that adding extra features to the model was causing overfitting, and the best results are found by keeping only the country and year as predictors.

If you changed the strategy for missing values, what did you change it to and why?

For missing life expectancy information, linear imputation seems like a good method to complete the data. This puts a value in-between two adjoining data points, which makes sense (big variations in-between two years in life expectancy does not happen).

For the GDP, the data concerns GDP growth. On average, it can be thought that this stays relatively close to 0. Linear interpolation was attempted but did not significantly improve the model.

Strategies for HIV rates, income per capita, and internet use were explained above.

What data or features would you have liked to include but were not able to acquire?

Additional features that would have been interesting to include, but that I did not, include access to healthcare, vaccination data, education level, infant and male and female mortality rates. Indeed, I expect that an increase in access to healthcare, as well as in vaccination rates and education level in a country would lead to an increase in the mean life expectancy. Likewise, a decrease in the mortality rates should correlate to a higher life expectancy. However, I did not try to include these features. Based on the results I got by including the features that I talked about in this report, I also expect that adding these extra features would not improve the model much more.