Assignment 3

Shivakanth Thudi

12/01/2016

Analysis

Part 1 - Algorithms Considered

I considered linear regression, decision trees, random forests, gradient boosting machines, and extra-trees regressors. I also built a validation set for predicting life expectancy for years outside the ranges of the training data provided, specifically for the years 1960 and 2011 to 2015. This data was not used to train the model; it was only used to validate the model's performance.

The training MSE and validation set MSE are as below:

Model	Training Set MSE	Validation Set MSE
Linear Regression	0.875	1.530
Decision Tree Regressor	8.15e-11	1.747
Random Forests	0.073	1.099
Gradient Boosting Machines	0.015	0.822
Extra Trees Regressor	9.975e-12	**0.766**

The extra trees regressor performed the best on the validation set, so I picked this algorithm for my final model.

For all the regressors except linear regression, I found the optimal number of trees to use by averaging validation set MSE over 10 runs. For the extra trees regressor, the optimal number of trees was 40, so I used this number in my final implementation.

Part 2 - Data added for features

For my baseline model, the only features included were the Year and the Country Name. For the final model, I also included the following as additional features:

- Birth Rates
- Male and Female Mortality Rates
- Infant and Under-5 Mortality Rates
- GNI per capita
- HIV rates
- Internet users per 100 people

These features were each added to the baseline model, and were found to decrease the validation set MSE. These results, combined with my research and intuition, led me to add these features to my model.

The following table illustrates some of the changes in MSE after addition of the above features:

Algorithm	Baseline Model MSE on validation set	Augmented Model MSE on validation set
Linear Regression	16.96	1.530
Random Forests	1.584	1.099
Gradient Boosting Machines	1.017	0.822

Part 3 - Strategy for Missing Values

- **Response** - For missing values in the target data, i.e., life expectancy, I imputed the values using linear interpolation.
- **Features** - I ended up using the same strategy of linear interpolation for the features in my model, since this seemed to improved the model's performance the most on the validation set. The other strategies I used (replacing missing values with 0, spline interpolation) did not improve the model's performance in comparison with linear interpolation.

Part 4 - Additional features that I would have liked to include

Some of the other features I would have liked to include are:

- Total expenditure on health per capita
- Pollution estimates

I was unable to acquire accurate estimates for the above features, and data was mostly unavailable for the years we considered. However, I believe these features would help predict life expectancy more accurately.