World Bank Population Data Analysis

Harinath, Ashish Kumar, Sambandh

Abstract

Activity 3 is based on the World Bank Data from kaggle, which aggregates the population of various countries, along with fertility rate and life expectancy, from 1960 to 2016. The goal of this activity is to explore a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and limit the fragility of a statistical model.

I. Introduction

To predict the population of a single country taking into consideration 5 other countries which can be modelled together to build a linear regression model so as to get the least Mean Squared Error(MSE) over the testing dataset.

1) From the dataset country-population.csv, we go on to eliminate the extraneous columns which have missing values and the incomplete rows.
2) Then, we use this data spanning for a period of 40 years i.e. the population estimate of 258 countries from 1960-2000 to build a linear regression model that could predict the population of a country over the test data.
3) **Training Data:** The population data of 258 countries spanning over a period of 40 years i.e. from 1960 to 2000.
4) **Testing Data:** The population data of 258 countries spanning over a period of 17 years i.e. from 2000 to 2017.

A. Procedure

1) The dataset population-training-kaggle.csv has about 40 rows X 259 columns where the columns represent the country names and the rows represent the years from 1960 to 2000.
2) Predicting the population of a single country based on the population estimate of 259 countries is a daunting task as such a model would lead to overfitting.
3) That is why, we go for dimensionality reduction techniques such as Ridge regression and LASSO.

II. Least Squares and Regularization

1) **LEAST SQUARES:** The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems. This technique is used extensively in the context of data fitting; the best fit in the least-squares sense minimizes the sum of squared residuals.
2) But in our case since we have$p(predictors) > N(samples)$, we need to perform Variable selection.
3) **TIKHONOV REGULARIZATION:** In the current context, a regularized version to the least squares solution is highly desirable. Ridge regression, or Tikhonov regularization, adds the constraint that the L_2-norm of the parameter vector remains no greater than a given value.

$$(1/(2 * N - samples)) * ||y - Xw||_2^2 + alpha * ||w||_2 \tag{1}$$

4) **LASSO:** An alternative regularized version of least squares is Lasso (least absolute shrinkage and selection operator), which uses the constraint that the L_1-norm of the parameter vector be no greater than a given value. The latter constraint, which we will focus on, favors sparsity in the solution. Using Lasso regularization to perform this

$$(1/(2 * N - samples)) * ||y - Xw||_2^2 + alpha * ||w||_1 \tag{2}$$

III. Code Flow:

1) Data processing: to find missing values or find anything other than the desired format.
2) Standardizing the data: Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning algorithms (e.g., support vector machines, logistic regression, and artificial neural networks). The general method of calculation is to determine the distribution mean and standard deviation for each feature. Next we subtract the mean from each feature. Then we divide the values (mean is already subtracted) of each feature by its standard deviation. Where x is the original feature vector, $\bar{x}$ is the mean of that feature vector, and σ is its standard deviation. $x' = \frac{x-\bar{x}}{\sigma}$
3) Implemented LASSO using existing ScikitLearn module.
4) A While loop to iterate over all the 258 coutries as Y and run lasso over wide range of alphas to pick model which has $\hat{C}_0(year) = \sum_i \alpha_i C_i(year)$

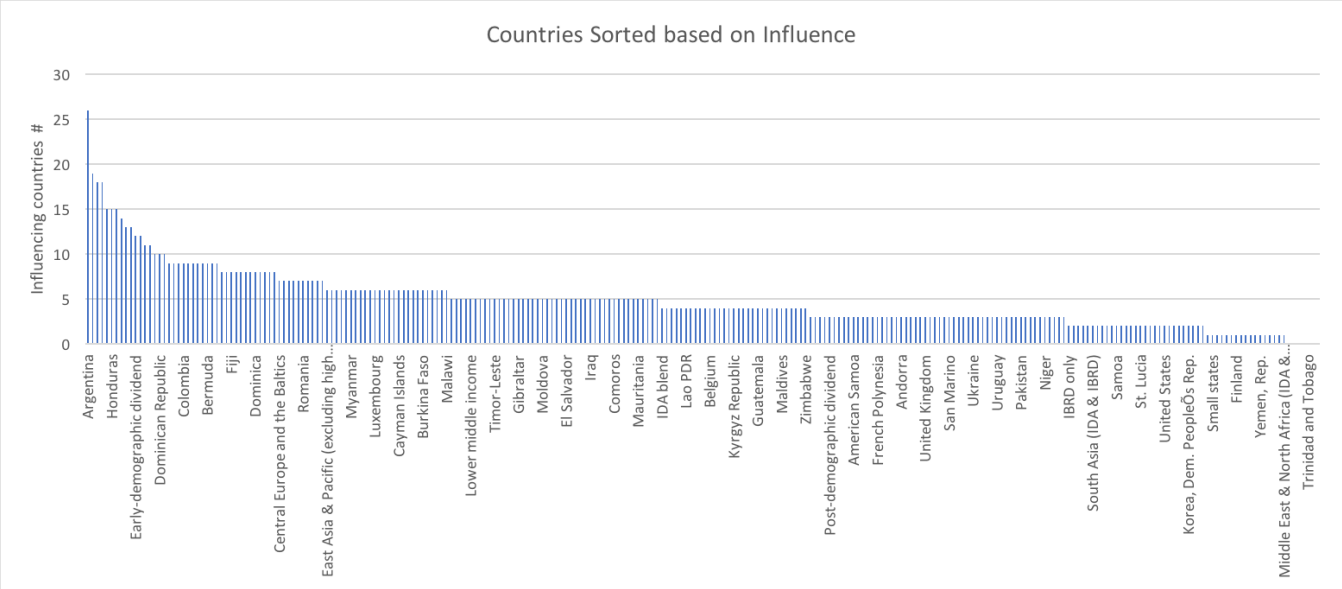

Fig. 1: Most influencing country

IV. Conclusion

1) Using the training data of 40years and model trained using atmost 5 other vectors, we predicted the population of all 258 countries on test data.
2) Outputs of the code: Prediction and coefficient matrix are shared in GitHub.
3) In addition to that we tried plotting the data in Excel and Gephi.
4) We defined a metric Influencing Countries : if a country (X) appears in model of 20 different countries than this metric value is 20. So we picked number of different models in which a country has a relation with.
5) You can see the figure "Most influencing Country" , Argentina appeared on 26 different models and hence affected most countries population.
6) In addition to above observation : we tried to plot the whole coefficient matrix as Adjacent matrix into the Graph , nothing concrete could be interpreted through that. As expected that graph was very sparse, so it didnt have a dense look at all.
7) So, idea is that we will group the countries into different regions and plot the adjacent matrix between that. We will take a look at them as different regions, namely: SubSaharanAfrica, LatinAmerica Carribean etc.
8) This way of representation helped us visualize and understand the data better. We could see that grpah is now dense meaning the countries in same region have good amount of edges(relation / coefficient) between them. This makes real sense too, since there is always high change of immigrants migration between neighbouring countries.
9) Another hypothesis is that, we see countries like Argentina (region: Latin AmericaCarribean) , Arab world , Congo (Sub saharan african) which has maximum edges to countries in their regions beacuse their economic conditions are better in general when compared to other countries in that region!

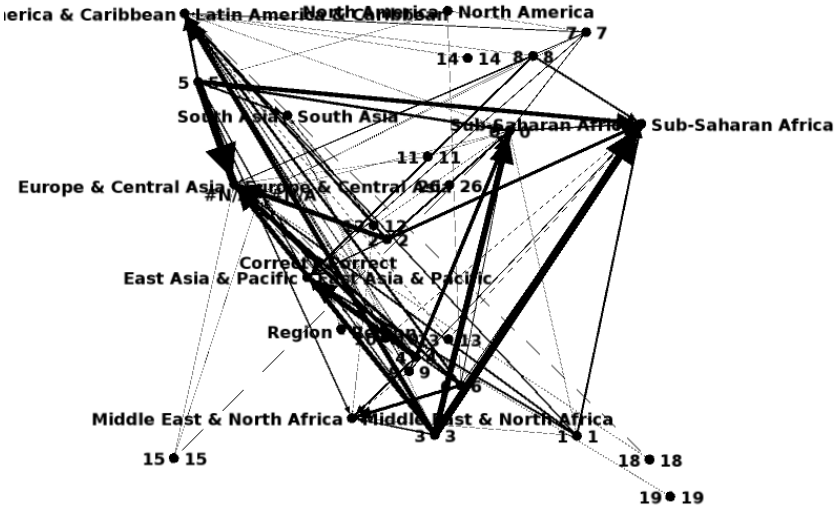

Fig. 2: Connections between regions

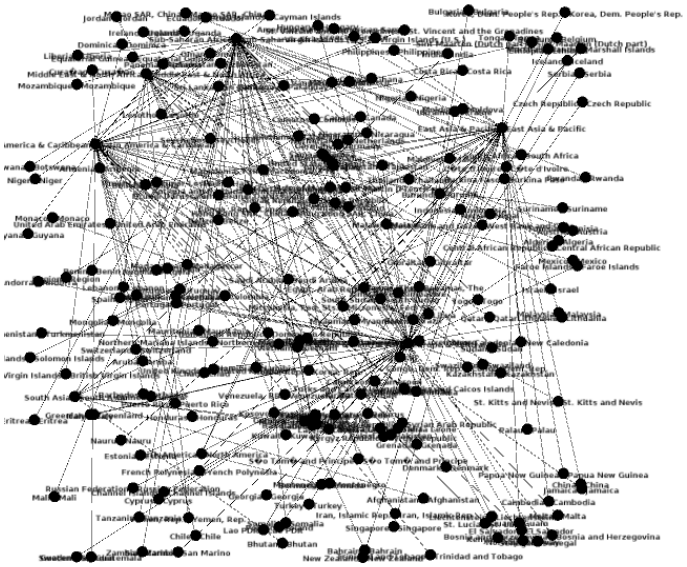

Fig. 3: All countries connected through an edge if coefficient values is non zero — High Sparsity

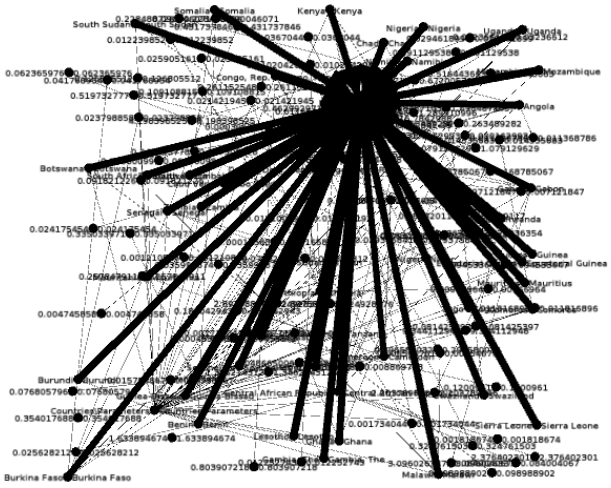

Fig. 4: Sub Saharan Region Map : Dense compared to the map over all regions