Script for the Machine Learning Predictions

Adrien Ratsimbaharison

March 16, 2019

## The six steps in the machine learning predictions:

1. Defining the problem
2. Reading and exploring the Data
3. Preparing the data: data randomization and partition
4. Predicting with classification
5. Predicting with regression
6. Conclusion

### Step 1: Defining the problem

When performing machine learning techniques, it is always a good idea to define exactly What you are trying to predict. In our case, we are trying to predict:

1. the classification of countries either as “stable” (with stability scores greater than 0) or as “unstable” (with stability scores less than 0), and
2. the stability scores of countries based on some predictors.

In terms of machine learning, we are dealing in this case with a multivariate supervised machine learning problem in which we have to predict a binary and multi-class outcome (thus we will use classification techniques) and a numeric outcome (thus we will use a mutliple regression technique).

### Step 2: Reading and exploring the Data

Our data were downloaded from different sources and merged in a single data frame (See script on downloading quantitative data for the information on how the data were collected and manipulated).

We just need here to read the data into r and explore the features.

Before preparing the data for the modeling, we need to take a look at the target variables: stabilityDummy and stability.

We need also to take a look at the summary statistics of the other variables of interest, paying attention to the existence of NAs, outliers, and zero variances.

From the summary statistics of the other variables, we can see that, among the Polity IV variables, there are some extreme values of -66 (for interruption periods), -77 (for interregnum), and -88 (for transition periods), which are known as “standardized authority codes” (See Marshall, M. G., Gurr, T. R., & Jaggers, K., 2018). Since there is no commonly agreed procedure on how to handle these extreme values among scholars (See Plümper, T., & Neumayer, E., 2010), we decided to remove them from our data, because they may falsify our results.

### Step 3: Preparing the data: data randomization and feature selection

The data were arranged in a country-year format. Therefore, it is a good idea to randoming the rows before partitioning the dataset

### Step 4: Predicting with classification

### Step 5: Predicting with regression

### Step 6: Conclusion