MSAN 621 Introduction to Machine Learning Fall 2016 Module 2

Assignment 3

There are two portions to this assignment: a practical portion and an analysis portion. For the practical portion, you will build a model using one of the machine learning algorithms covered in class. Once complete, you will answer the questions below. You will submit your implementation code for the practical portion and your answers (as a text file or document in Word, OpenOffice or PDF) for the analysis portion.

Practical Portion

For the practical portion, you will use one or more of the machine learning algorithms covered in class to predict life expectancy for the "average" person born in a certain year in one or more countries. There are two files are attached to this assignment. The files share a (common) set of labels: countries appear in rows and years from 1961 - 2010 appear in columns. The first file, `life expectancy by country and year.csv`, contains the targets for your model. The second file, `GDP by country and year.csv` contains a feature for your model, should you choose to use it. Note that there are several missing values from both files.

A baseline model produces coefficients [0.30798363 -0.03583604] with two features (year and GDP) used to predict life expectancy. In addition to these features, the baseline model makes the following assumptions for missing values:

- Missing GDP values are filled with 0.0
- Missing life expectancy targets are assumed to have values which increase or decrease in a linear fashion; such missing values may occur between known values, prior to known values or subsequent to known values

You may choose to reproduce this baseline model but you are under no obligation to do so.

You may [read: should] implement a model which takes this data and produces hypotheses about the life expectancies of people in the various countries. Your model may employ any (single) machine learning algorithm or approach we discussed in class, but it behoves you to choose the best one given the data provided. With the exception of additional life expectancy data or predictions of life expectancy, you may add any external data to your model with the aim of improving the prediction accuracy.

You may expect the two files discussed above to be available to your implementation in the same directory as your implementation file(s). Two parameters will be provided to your implementation. In order, the parameters are 1) The name of the input file and 2) The name of the output file. Your implementation must be callable from command line, as follows:

```
python my_a3_script.py input.csv output.txt
```

The input file is a CSV-formatted file, with the following fields on each line:

- Country Name, which is an exact value match to an entry in the first column of the `life expectancy by country and year.csv` file
- Year of prediction, which is between 1950 - 2016
- GDP for the country and year

An example of a row from the input file is as follows:

```
Vietnam, 1955, 1.99483712943539
```

The output file is the prediction vector of your implementation for the input file. It should contain one life expectancy prediction per row from the input file, and the predictions must appear in the same order as the input file. (Row 1 in the output file must be the prediction for row 1 in the input file.)

If you choose to add additional features to your model, you must provide the data for those features, which should be in the form of files in a format of your choosing. Note that any added data or features must be generatable for the years 1950 - 2016 in order to support your prediction model.

Analysis Portion

Based on the results of your implementation above, answer the following questions:

1. 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?
2. If you added data for features, what are the sources of the data and why did you add the data from these sources?
3. If you changed the strategy for missing values, what did you change it to and why?
4. What data or features would you have liked to include but were not able to acquire?

Grading

The implementation portion of this assignment is worth 80% of the grade, and the analysis portion the remaining 20%, evenly divided among the questions. The implementation portion is graded as follows:

- 60% = Proper implementation of the prediction model with three features (country, year, GDP) as provided by the input file
- 40% = Performance incentive, with highest-performing model above baseline receiving full 40%

Late assignments will receive 0%.