CIS520 - Group 10

Amit Gupta, Monal Garg, Aashish Jain, Moksh Jawa
akgupta@seas, mgarg@sas, aashj99@seas, moksh@seas

Project Proposal

March, 2019

1 Introduction

1.1 Problem Motivation

Unfortunately, it is a sad reality that many individuals each year commit suicide. We in the U.S. are slowly being aware of this issue, but we wanted to understand suicide rates internationally. Especially, we wanted to look at these rates as they related to common features of each country, such as GDP, unemployment, and alcohol consumption. At a high-level we wanted to draw out any meaningful relationship between these macro features and suicide rates.

1.2 Data Set and Source

Our project grew from an intriguing World Health Organization data set will well over 35 years of suicide rates by country broken up by demographics. [1]. In an effort to extend existing analysis though, we actually sources global data from a variety of sources. This included a FiveThirtyEight data set on alcohol consumption [2]. We also incorporates a World Bank Data Set on Youth Unemployment [3] and another from the World Bank on population, fertility, and life expectancy [4].

1.3 Summary Statistics

TODO(nicely formatted table)

2 Related Work

2.1 Related to Models

Among the papers that analzyed economic factors, a regression model seemed to be the popular choice of method. For example, Pandey and Kaur paper used a Auto-Regressive Distributed Lag (ARDL) model [5].

Most of the articles concerning the relationship of suicide rate and alcohol consumption related trends from a biological perspective utilized relatively simple statistical analysis [6]. Whereas medical research that involved more complex forms of data such as fMRI brain scans utilized algorithms such as Gaussian Naive Bayes [7].

2.2 Related to Problem

During our research, we came across multiple research papers that analyze the correlation between suicide rates and either economical, behavioral, medical, or psychological factors. For example, the Pandey and Kaur paper investigates suicidal trend and economic determinants in an Indian population[8]. Whereas articles published in medical journals explored suicide with alochol/drug consumption, psychotic behavior, brain scans, etc [9].

However, most papers tend to focus on either the economic or behavioral/medical factors. As suicide is likely governed by multiple factors, we seek to investigate this problem through a more holistic approach by simultaneously analyzing both medical, behavioral, and economic factors. Through this application of machine learning, we hope to gain insight on the relative role each factor plays in suicide and about the interaction of these factors.

Moreover, most papers perform a time series analysis. We provide a different approach by analyzing the data by country. Papers that do focus on certain regions usually provide analysis on one country or a specific region, whereas we will analyze global statistics.

[1] https://www.kaggle.com/szamil/who-suicide-statistics
[2] https://www.kaggle.com/fivethirtyeight/fivethirtyeight-alcohol-consumption-dataset
[3] https://www.kaggle.com/sovannt/world-bank-youth-unemployment
[4] https://www.kaggle.com/gemartin/world-bank-data-1960-to-2016
[5] https://crawford.anu.edu.au/acde/asarc/pdf/papers/2009/WP2009/_08.pdf
[6] https://www.ncbi.nlm.nih.gov/pubmed/9229027
[7] https://nocklab.fas.harvard.edu/files/nocklab/files/just_2017_machlearn_suicide_emotion_youth.pdf
[8] https://crawford.anu.edu.au/acde/asarc/pdf/papers/2009/WP2009/_08.pdf
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2872355/

3 Problem Formulation

3.1 Data Processing Needed

Our data came in a variety of comma-separated formats. A lot of countries were labelled in differing manner and years were often a column and other times a point in a row. We used Excel Pivot Tables to make these more easily parsable and standardized. Later, blanks and other issues with inconsistencies are removed with Python to get an aggregated set of data. Other pre-processing includes removing blanks, reconciling some country names such as (Trinidad & Tobago vs. Trinidad and Tobago).

3.2 Inputs, Outputs

Inputs are a variety of real-valued data points tied to the country. They are described above, but it remains to see which we might have to throw out due to excessive blanks. Most are also measured for a specific year, however some are just duplicated across all years (e.g. Alcohol Consumption). Outputs are a model, which given a countries characteristic features, can predict suicide rates. We have suicide rates by gender (as well as age demographics), so we may also seek a model which predicts a rate for male suicides and a rate for female suicides.

3.3 Performance Measures

Given the regression nature of this problem, we believe a squared loss model is appropriate. However, during the course of our work, we may seek alternate loss models such as a proxy for TPR and TNR through some acceptable error window.

4 Solution Methods

4.1 Algorithms Planned For Use

We will start with least squares regression in unregularized, L_1 regularized, and L_2 regularized forms. We are not certain whether any of our features will be overly correlated so the unregularized least squares regression may fail. We also seek to explore Nearest Neighbor methods, but not with vanilla Euclidean distance, but with feature weighted scaling on the distances. Lastly, we will consider using decision stubs as features and feeding the transformed input space into another learning model, perhaps Neural Nets.

4.2 Justification

As a regression problem, squared loss makes sense as a first choice. Furthermore, we want to limit ourselves to models which actually inform us about the relative importance of features. As such, vanilla nearest neighbors is not informative in the context of our problem. However, determining similarity in which features best informs Nearest Neighbors makes sense given our problem motivation.

5 Plan of Work

5.1 Team Member Tasks

Monal: Past Work, Solution Approach, Machine Learning Algorithms, Visualizations, Report

Amit: Problem Formulation, Algorithms Plan for Use, Machine Learning Algorithms, Visualizations, Report

Ash: Data Cleaning, Data Stitching, Machine Learning Algorithms, Visualizations, Report

Moksh: Data Cleaning, Data Stitching, Machine Learning Algorithms, Visualization, Report

5.2 Pre-Milestone Meeting Goals

Our goal before our milestone meeting is have the data completed clean and visualized. Additionally, we plan to have k-Nearest Neighbors and Least Squares Regression implemented, visualized, and results recorded. Note that our k-Nearest Neighbor implementation will have weighted scaling based on features rather than Euclidean distance.

5.3 Timelines

After our milestone meeting, we will have exactly 3 weeks. The first week will involve implementing our last approach of decision stubs as features and feeding them into neural nets. The second week will involve finetuning and evaluating our models. The last week will be for writing up the report.

6 Acknowledgements

N/A

7 References

In footnotes

8 Appendix

None yet.