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

TODO

2.2 Related to Problem

TODO

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

[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

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

TODO

5.2 Timelines

TODO

5.3 Pre-Milestone Meeting Goals

TODO(detailed)

6 Acknowledgements

TODO

7 References

TODO

8 Appendix

TODO