Research Update
Machine Learning: Final Project
Bobbie van Gorp & Beau Furnée
12-12-2017

Changes to Research Proposal

The initial idea was to study the happiness scores of all the countries of 2017, but we soon realised that there were two problems with this. First, since 2017 is not over yet, not all the data we want is available yet. Second, because of this same reason, the happiness scores were not yet completely calculated, and are still subjectable to change. We solved these to problem by choosing 2015 to be the year we study.
Another problem we ran into, was that our dataset seemed to contain all the data we needed, but this was not completely true. In our dataset, it did not display the actual values of features like *GDP* and *Life Expectancy*, but "The extent to which {the feature} contributed to the calculation of the happiness score". Therefore, we ended up completing our dataset ourselves, through finding the data we needed in additional datasets (see 'Sources').

Algorithms

To begin with, we would like to use Linear Regression, since this is an algorithm were very comfortable with. Since we expect our dataset to be nonlinear, we think working with a Support Vector Machine will result in better regression and also provides a good transition into classification.
The labels in the dataset range from 1 to 10, which grants us the opportunity to classify the happiness scores into 10 clusters (1-2, 2-3, 3-4, … , 8-9, 9-10). Using either a Decision Tree or a Random Forest algorithm gives us a whole new angle of approach and might result in critical new insights to better understand our dataset. Therefore, we would also like to give these algorithms a try.

Evaluation

To compare the algorithms, we will feed them multiple new real examples from the dataset and see how they perform. Also, it will be interesting to see which algorithms prefer which features and try to figure out why this would be the case. We can accomplish this by feeding the algorithm 'fake' examples with exaggerated features and see how that influences the happiness scores.

Progress

Thus far, we have successfully completed our dataset using multiple new sources and cross-checked them. Besides filling in missing values, we have added an additional feature to our dataset that we thought will results in interesting results. Because we are going to use classification as well, we thought it would be interesting to add a binary features. We chose to add whether a country is a member of the OECD as that binary feature, since this creates a nice division between first world countries and the rest.
Also, we have chosen which algorithms we will be using based and their individual relations to our dataset and their respectively inherent differences such that we will achieve the optimal results and the most interesting comparisons.

Sources

https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

https://data.worldbank.org/indicator/SP.DYN.LE00.IN
https://www.kaggle.com/unsdsn/world-happiness/data
https://www.cia.gov/library/publications/the-world-factbook/rankorder/2129rank.html
https://knoema.com/atlas/
https://www.ncbi.nlm.nih.gov/pubmed/27437333
https://tradingeconomics.com/
https://data.worldbank.org/indicator/SE.ADT.LITR.ZS