Research Proposal
Machine Learning: Final Project
Bobbie van Gorp & Beau Furnée
05-12-2017

Problem

For our research, we will analyze the happiness of the population of different countries. We will try to predict happiness scores and also discover which factors are most influential on the scores. Besides analyzing the dataset, we will also compare algorithms and their outcomes using this same dataset. Our final report will thus consist of an analysis of the dataset and an evaluation of the algorithms applied.

Dataset

The dataset we will be using can be found through https://www.kaggle.com/unsdsn/world-happiness. It is divided into 3 years which consist of 155 examples, portraying countries and their happiness scores. Besides the happiness scores, each country also has the following attributes:

- economic production;
- social support;
- life expectancy;
- freedom;
- absence of corruption;
- generosity.

All the attributes are numerical. The happiness score is based on the main life evaluation question in a poll from the Gallup World Poll and so is not directly calculated from any of the features in the dataset.

Method

In order to estimate the happiness score of a country based on the dataset, we will be applying three different algorithms on the same dataset. The algorithms we chose are:

- Linear Regression;
- Decision Tree;
- K Nearest Neighbors.

These will be trained using a part of the dataset to come up with a formula which will be tested using a different part, to see whether it is accurate in predicting the happiness scores. If this indeed is the case, we will use the trained algorithms to analyze the attributes of the datasets and see which cause the biggest changes in the happiness scores. We can accomplish this by feeding the algorithm ‘fake’ examples with exaggerated features and see how that influences the happiness scores.

Evaluation

Finally, we will compare the algorithms and see how the results differ. We will do this by feeding them multiple actual examples from the test set 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.