Predicting Online News Popularity

Michael Steiner

Statement of the project and summary of the work completed (temporary section)

Summary of project: we will predict popularity of news articles based on a subset of 60 features found in the data set provided by UCI machine learning: https://archive.ics.uci.edu/ml/machine-learning-databases/00332/

This project has three distinct parts. In the first part, we will implement the six best performing machine learning algorithms employed by Ren et al. **in Predicting and Evaluating the Popularity of Online News**. The algorithms are linear regression, logistic regression, SVM (polynomial kernel), random forest, k-nearest neighbors, and SVM (linear kernel).

The second part will consist of exploratory work. The goal will be to implement a machine learning algorithm with has an accuracy of at least 70% which was the stated goal for Ren et al. in their future work section.

In the third part of the project, we will employ Mapper (http://danifold.net/mapper/introduction.html) in an attempt to see if topological data analysis can give us any additional insights.

Summary of work completed:

The data has been normalized, and NA values in the data have been handled. We have performed linear regression, SVM (linear kernel), and SVM (polynomial kernel) on the data set.

Outline of paper:

1 Introduction

Problem statement and Hypothesis

2 Materials and Methods

Description of data set

Where it was obtained

Prepossessing

Details of the modeling process

3 Results

4 Conclusions

What I learned

Challenges and successes

Conclusions and key learnings

5 Future work

Possible extensions or business applications

I.