CS 7280 Project Proposal

Gabriel Bakiewicz

Darshan Patel

Problem Description

The Online News Popularity Data Set has a set of features extracted from the articles published by Mashable in a period of two years. The primary goal is to predict the number of shares in social networks. Along with it, we would like to convert this regression problem into a classification problem by grouping the number of shares into different classes.

Data Summary

There are 61 attributes in the dataset. In them, 58 attributes are predictive, 2 attributes are non-predictive and a goal attribute. There are 6 attributes that represents one hot encoding of categorical attribute “Data Channel” and 8 attributes that represents one hot encoding of categorical attribute “Day of Week”. Along with it, there are 18 attributes which are different summary (Avg, Min, Max) values of 6 underlined attributes such as polarity of positive/negative words.

Given that we have large number of attributes; we would do some exploratory data analysis. We will check zero variance features and if there are missing value then will try to replace them with mean value of an attributes. Apart from that, will look for the pairwise highly correlated features.

Methods

We will use Linear Regression for predicting the number of shares and Multinomial Logistic Regression to predict a class of a target. For the features selection,

To reduce the over fitting, we will use the validation dataset.

Reference:

K. Fernandes, P. Vinagre and P. Cortez. A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News. Proceedings of the 17th EPIA 2015 - Portuguese Conference on Artificial Intelligence, September, Coimbra, Portugal