Tem Yousafi

Professor Linda Sellie

Machine Learning

25 April 2018

Write-up

# Introduction

For my project, I attempted to formulate three different machine learning models for the purpose of measuring the sentiment of tweets. I used k-means clustering, linear regression, and logistic regression.

# Background

Firstly, I developed a method of offering a rough evaluation of the sentiment of a tweet (i.e. how “positive” or “negative” the tweet is) using SentiWordNet and an enormous dataset of tweets taken from Kaggle, where tweets were given scorings of 0 for negativity and 4 for positivity.

My method involved tokenizing each word in the tweet, and searching SentiWordNet for each word’s individual “sentiment,” meaning its positive/negative scoring based on the SentiWordNet library. I then combined the scoring of each word to come up with an overall net-sentiment of the tweet: a rough method that has much area for sophistication, but one that suited my purposes effectively.

# K-means Clustering

The first method of supervised learning was the k-means clustering algorithm. I sorted a portion of the dataset (3000-6000 tweets) into two clusters. Essentially, the hope was that the algorithm would determine that the most obvious way to categorize the tweets was based on positivity/negativity, which it did.

The algorithm was far more accurate at categorizing positive tweets (6.4 failure rate) than negative tweets (65% failure rate). However, I determined that a fair method of regularizing the data was by eliminating relatively “neutral” tweets from the dataset (tweets only slightly negative or slightly positive, specifically between the sentiment-values of -0.1 and 0.1), in order to see how well the model could accurately cluster-together “extreme” tweets (i.e. tweets which were extremely negative or extremely positive). This regularization was shockingly fruitful. The model still categorized positive tweets together far more accurately (0.15% failure rate) than negative tweets (4.1% failure rate), but both failure rates were reduced drastically.

# Linear Regression

Because single linear regression requires a quantitative rather than qualitative dependent variable, I scaled each tweet’s polarity by the length of the tweet (number of words), my rationale being that a tweet which manages to stay positive over the course of a higher word count can be categorized as being more positive than a positive tweet which has far fewer words.

Such a model has virtually no linear relationship. The b0 value was determined to be 41.3 and the b1 value was 65.8, but the distribution of the manipulated data seemed to follow a normal distribution far more than a linear relationship, interestingly enough. The distribution can be seen on slide 7 of the presentation, and should be noted for future considerations.

I regularized the data once more by using a different approach: Rather than scaling polarity with tweet-length, I calculated a “polarity per word” instead, my rationale being that an overall-positive tweet will have, in a sense, an “average” word-sentiment that is positive. That is, if a tweet is positive, this overarching positivity should, on average, correspond to a smaller amount of positivity found in each word.

Such a model still has a linear relationship with much to be desired. The b0 value was 0.356 and the b1 was 1.67, but the sentiment my model evaluated had virtually no linear relationship with the “polarity per word” of tweets.

# Logistic Regression

I attempted to find the probability of a tweet’s polarity being positive given a positive evaluation by my own method of determining sentiment.

I used a method of trial-and-error for the maximum likelihood algorithm in determining the b0 and b1 values. I eventually found that a b0 of -395 and a b1 of 99 were suitable coefficients for my purposes. Without manipulating any data, I attempted to categorize tweets using this model of logistic regression, but only had a success rate of 7.5%.

However, I determined that using a similar regularization method as in my earlier k-means clustering model could prove fruitful. I attempted to remove the “noise” of relatively “neutral” tweets (by removing data points with a determined-sentiment of less than 2), and found that when my model categorized the leftover “extreme” tweets, it did so with far more accuracy: it had a success rate of approximately 60%.

Works Cited

<https://www.kaggle.com/kazanova/sentiment140/data>

<http://sentiwordnet.isti.cnr.it/>

<http://ptrckprry.com/course/ssd/data/positive-words.txt>

<http://ptrckprry.com/course/ssd/data/negative-words.txt>