Influencers in Online Social Network

# Introduction

Many entrepreneurs are using social media analytics to improve their marketing techniques and to expand their client base. One of the key components of social media analytics is utilizing patterns in online social networks, mainly using machine learning algorithms for predictive analysis. The following will explore an example of one of the uses of machine learning, namely pairwise preference learning which is used to predict total order (ranking). The goal of this project is to predict the human judgement of who are the influential members of an online social network. We will be aiming to derive an algorithm from the training set and apply it to the test set to predict which individual would be considered more influential by a human being. To solve this problem we will use logistic regression in R.

# Literature Review

1. Fürnkranz, Johannes and Eyke Hüllermeier. “Preference Learning: An Introduction,” in *Preference Learning*, 1-17. Berlin: Springer-Valen, 2010.

In Fürnkranz and Hüllermeier’s chapter introducing preference learning, they outline three different cases: label ranking, instance ranking and object ranking. For the purpose of this project, we are concerned with object ranking. In this case, ‘objects’ are ranked using pairwise preferences (or binary preferences), meaning that the ranking includes the objects that are chosen over their neighbouring options. In machine learning, one can use “a set of objects described in terms of multiple attributes” represented by these binary preferences to predict future preferences.

1. Li, Hang. “Data Labeling,” in *A Short Introduction to Learning to Rank*, 2. Tokyo: IEICE, 2010.

In this article, Hang discusses the theory and methods behind ranking the relevance of web pages. Although this is not the same problem we are exploring in this project; however, it provides insight on how the training dataset was created. One of the methods in created page ranking training sets is to start by randomly selecting queries from a query log. Then imputing those queries into various search engines and noting the multiple documents that are associated with each query. Finally, human judges rank the documents based on relevance using an ordinal scale.

1. Benevenuto, Fabrício and Meeyong Cha, Krishna P. Gummadi, Hamed Haddadi. “Measuring User Influence in Twitter: The Million Follower Fallacy,” in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 10-17.AAAI, 2010.

Benevenuto and his colleagues examine the role of influence in sociology, focusing on influence in the online social networks of Twitter. In this article, they highlight the importance of understanding online influence when it comes to marketing and persuasion. They describe these influencers as “informed, respected and well-connected”; moreover, they present the argument that now people are just as influenced by their peers rather than specific influential people. They consider three ways of measuring ones influence on Twitter: number of followers of a user, number of retweets of a user, and number of mentions container a user’s name.

1. Rodríguez, Germán. “Logit Models for Binary Data,” in *Lecture Notes on Generalized Linear Models*. New Jersey: Princeton University, 2007.

In this chapter, Rodríguez gives an example of how to build a regression model for dichotomous data, which is called logistic regression. It discusses the theory and formulas behind logistic regression, where the dependent variable is of the type ‘success or failure’ (derived from the Bernoulli trial) and the outcome is based on independent variables or attributes.

# Dataset

The dataset has been obtained from [www.kaggle.com](http://www.kaggle.com), originally provided by the London-based company PeerIndex. PeerIndex administers services in social media analytics, including assigning a score to social media users to reflect their social capital. This dataset contains information about the relationship between pairs of Twitter users. Each record in the training set describes whether it is individual A or individual B that is considered more influential based on eleven pre-determined features from their Twitter activity, such as number of followers and number of retweets. Each record is labelled either as ‘1’ if A is the more influential user or ‘0’ if B is the more influential user.

The training data set has 23 attributes (including the column with the labels ‘1’ and ‘0’), 5500 records and no missing values. The attribute “Choice” is a categorical variable (specifically dichotomous), while the rest of the variables are of type ratio because they contain continuous, quantitative data (integers and real numbers) and the value ‘0’ represents the quantity zero. In addition, the observations are independent of each other and the dependent variable has mutually exclusive and exhaustive categories. Given these descriptions of the data set, the best predictive model would be a logistic regression model.

Since we have no information on how PeerIndex measures all the features of Twitter activity, it is difficult to make assumptions based on some of the attributes. Firstly, the attributes that are represented by real numbers because; for example, if Twitter user A has sent 1.111 retweets, we have no information to be able to interpret what 0.111 of a retweet is. Secondly, the data set includes three attributes for both user A and user B that are simply called “Network Feature” 1, 2 and 3. Without any information on PeerIndex’s methodology, we cannot speculate on what those features measure.

# Approach

## Step 1: Import dataset into R

Download dataset train.csv from [www.kaggle.com](http://www.kaggle.com) and import into R studio using the command read.csv(), then check for missing values using anyNA(). Change dependent variable type from integer to factor.

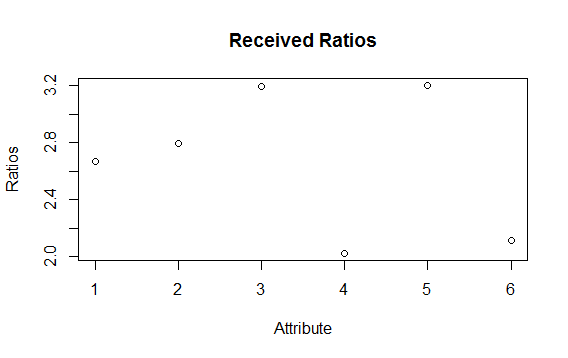
<https://github.com/lyndsayroach/Capstone-Crs--Project/blob/master/Step%201.R>

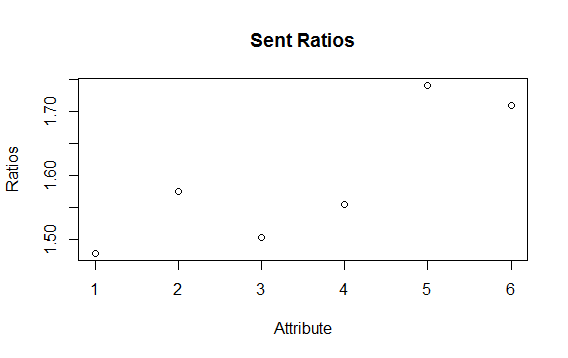
## Step 2: Explore dataset

Find patterns and compare records that are labelled ‘1’ and those that are labelled ‘0’. Preform summary statistics where applicable.

To begin exploring the dataset, we separate it into two subsets: the set labelled ‘1’ and the set labelled ‘0’. We will focus on the attributes follower count, following count, mentions received, mentions sent, retweets received, retweets sent, posts and listed count. I would predict that the Twitter actions done to the user are predictive of their influence. When we calculate the means of the follower count, following count, mentions received, mentions sent, retweets received and retweets sent, we can see that the individual that is more influential tends to have approximately 2.66 times higher mean of the number of Twitter actions done to their account and approximately 1.59 times higher mean of the number of Twitter actions they perform. For example, in the subset where individual A is more influential, individuals A have an average of 916888.4 followers and individuals B have an average of 343586.9 followers. Whereas, individuals A follow an average of 15721 accounts and individuals B follow an average of 10633.38 accounts. Also, we can assume that number of followers and number of posts may have a positive linear correlation with the attributes mentions received and retweets received. To check this we use Pearson’s correlation and observe combinations. For example, when individual A is perceived as more influential the Pearson correlation between A’s number of followers and A’s number of mentions is only 0.4228564. Most of the correlations calculated were pretty low, this could be because an individual does not need to follow a particular Twitter account to retweet them or to mention them in a tweet.

<https://github.com/lyndsayroach/Capstone-Crs--Project/blob/master/Step%202.R>





## Step 3: Build predictive model

Use the glm() command in R to build a logistic regression model than will help us predict the labels of the test dataset.

<https://github.com/lyndsayroach/Capstone-Crs--Project/blob/master/Step%203.R>

## Step 4: Concluding remarks

Concluding remarks depend on results.

# Results

Explain your results here. Consider that you need to communicate your results to executives in an organization. For example:

1. Insert tables and/or charts showing the results
2. Write description of the tables and charts, such that they show the usefulness for an organization
3. Identify the evaluation measures, such as accuracy, precision, recall, etc.

# Conclusions

Give a short summary (one to two paragraphs) of your analysis and conclude the discussion by defining the usefulness of your analysis.