Lyndsay Roach

February 15th 2016

CKME 136

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 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 attributes 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.

# Approach

## Step 1: Import dataset into R

Download dataset train.csv from [www.kaggle.com](http://www.kaggle.com) into R studio using the command read.csv() and check for missing values using anyNA().

## Step 2: Explore dataset

Use RSQLite and ggplot2 packages in R to explore data. Visualize characteristics of each record in training dataset, find patterns and compare those that are labelled ‘1’ and those that are labelled ‘0’. Preform summary statistics where applicable.

## 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.

## Step 4: Concluding remarks

Concluding remarks depend on results.