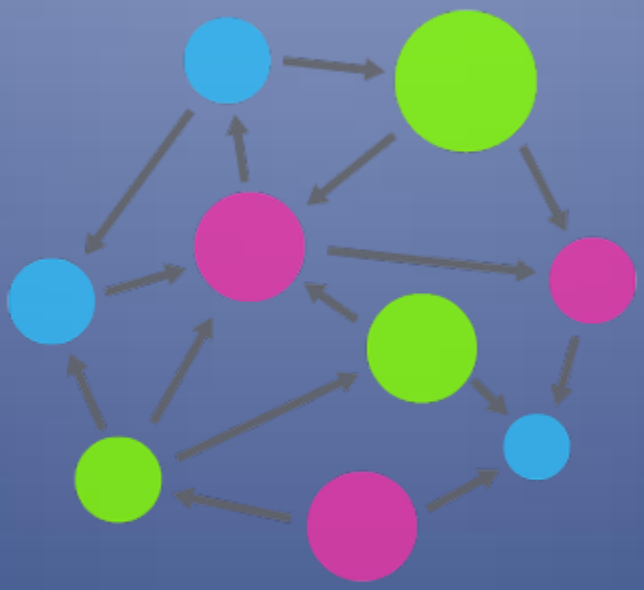


Influencers in Social Networks

Project Inspiration:

Influence has long been studied in the fields of sociology, communication, marketing, and political science. The notion of influence plays a vital role in how business operate and how a society function, see observation on how fashion spreads, and how people vote.

Studying influence patterns can help us better understand why certain trends or innovations are adopted faster than others and how we could help advertisers and marketers design more effective campaigns. Studying influence patterns, however, has been difficult. This is because such a study does not itself to readily available quantification, and essential components like human choices and the way your society functions cannot be reproduced within the confines of the lab.

Social media influence is much more than follower counts on Twitter or the volume of content shared on social platforms. It is about the ability to drive conversations and actions, and needs to be measured in context.

In this project I have tried to create a machine learning model which, for pairs of individuals, predicts the human judgment on who is more influential on twitter.

Something about Dataset

The dataset, provided comprises a standard, pair-wise preference learning task. Each data-point describes two individuals, A and B.

For each person, 11 pre-computed, non-negative numeric features based on twitter activity. These feature namely:

- Number of twitter followers
- Number of twitter following
- User listed count
- Number of times user received mentions
- Number of re-tweets user received
- Number of times user sent mentions
- Number of re-tweets user sent
- Number of user post
- Network feature_1
- Network feature_2
- Network feature_3

The binary label represents human judgment about which one of the individual is more influential.

A label '1' means A is more influential than B. '0' means B is more influential than A.

Number of training data-points	5500
Number of testing data-points	5952

Methodology for comparing user influence

Now we have data such that, for each data point, we have two users and a single label. Label defines who is more influential among the given two users.

This means, there is some difference between two users which makes one user more influential than other.

Hence we take the difference of feature values of two users.

Algorithm used for training predictor

I have used AdaBoost machine learning algorithm for training the predictor. Algorithmic details for the binary classification task can be found at

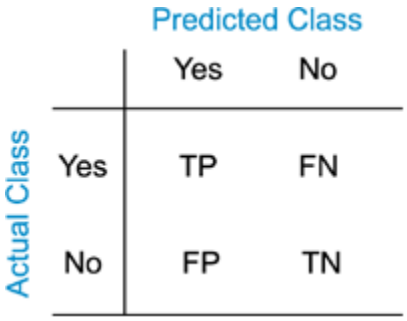
http://en.wikipedia.org/wiki/AdaBoost

Analysis:

I computed various parameters which help in analyzing the performance of the algorithm.

Confusion Matrix:

In the field of machine learning, a confusion matrix, also known as a contingency table or an error matrix, is a specific table layout that allows visualization of the performance of an algorithm.



TP = True Positive
FP = False Positive
TN = True Negative
FN = False Negative

After 57 rounds of boosting, following confusion matrix was generated.

		PREDICTED CLASS	
		A more influential	B more influential
ACTUAL CLASS	A more Influential	2828	224
	B more influential	151	2749

The correct guesses are located on diagonal of table.

The error rate for the given algorithm is,

(Number of misclassified test data-points) / (Total Number of test data-point)

= (151+224)/5952 = 0.0630040322581

Testing Error: = 0.0630040322581

Matthews Correlation Coefficient

The **Matthews correlation coefficient** is used in machine learning as a measure of the quality of binary (two-class) classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

After 57 rounds of boosting, the algorithm outputs following value for MCC:

MCC = 0.87425176318

A coefficient of +1 represents a perfect prediction. With MCC value of 0.88 for the implemented algorithm, it can be very much concluding that, the predictor decision are highly fair and accurate.

Accuracy:

The accuracy is the proportion of true results (both true positives and true negatives) in the population. It is a parameter of the test.

$$\mathrm{accuracy} = \frac{\mathrm{number\ of\ true\ positives} + \mathrm{number\ of\ true\ negatives}}{\mathrm{number\ of\ true\ positives} + \mathrm{false\ positives} + \mathrm{false\ negatives} + \mathrm{true\ negatives}}$$

Accuracy = 0.93699

F1 Score

In statistics, the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision *p* and the recall *r* of the test to compute the score.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

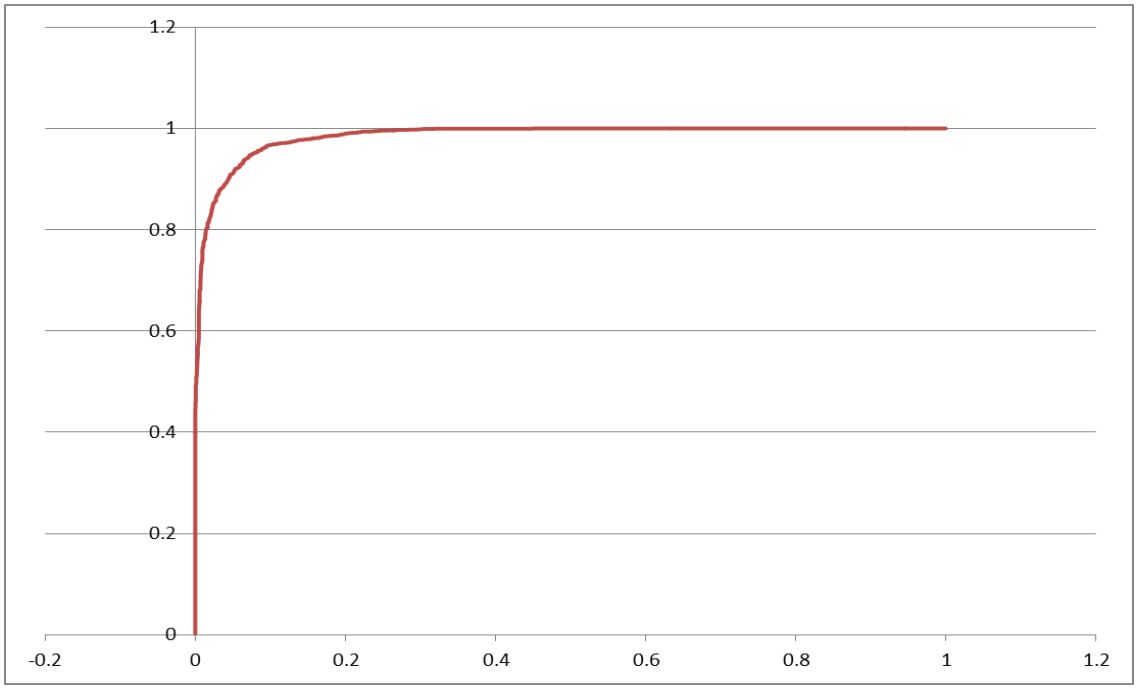
Precision = 0.95

Recall = 0.93

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F1 = 0.93782125684

Receiver Operating Characteristics (ROC) curve



X axis = False Positive Rate

Y axis = True Positive Rate

The ROC curve shows how the two rates change as the threshold changes. The leftmost point corresponds to classifying everything as the negative class, and the rightmost point corresponds to classifying everything in the positive class.

Area of Curve:

The area under the curve (simply the AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative').

After 57 rounds of boosting, AUC that classifier reached is **0.985023824694.**