Predicting News Sharing on Social Media

Joe Johnson
jjohnso3@stanford.edu

Noam Weinberger
noamw@stanford.edu

Summary

Goals:

- Predicting which articles are most popular would help news companies gain readers
- Estimation: Predict number of shares of news article on social media
- Classification: Predict whether article will be popular (shares above threshold value)

Methods:

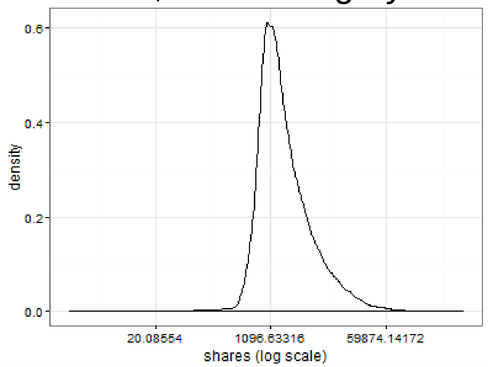
- Linear Regression, Logistic Regression, SVM, Random Forests, GDA

Results:

- Best Estimation: RMSE = 12,254 shares
- Best Classification: 66% accuracy

Dataset*

- 39,644 articles published from 2013-2014 on Mashable.com
- 60 article features, e.g. day of week, number of images, number of words, sentiment, news category



Feature Selection

- Removed columns that are linear combinations of other columns, e.g. day of week, news category
- Removed apparent errors
- Removed very recent articles
- Used Forward Selection
- Used PCA to reduce dimensionality
- Used Feature Scoring with Mutual Information

Top 5 Features

Feature	Coefficient
Avg Keyword	1.48
Max Keyword	-0.19
Entertainment Category	-1501.12
Links to Other Mashable Articles	-78.35
Title Length	125.81

Models

Linear Regression

$$y = \beta X + \varepsilon$$

Logistic Regression

$$y = g(\beta X + \varepsilon), g(z) = \frac{1}{1+\exp(-z)}$$

Gaussian Discriminant Analysis

$$y \sim Bernoulli(\varphi), x|y = 0 \sim N(\mu_0, \sigma_0)$$
$$y|x = 0 \sim N(\mu_1, \sigma_1)$$

SVM

$$max\ f(\alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n} y_i\, c_i(x_i, x_j)c_j$$

$$subject\ to \sum_{i=1}^{n} c_i y_i \ \ and\ 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all i}$$

Results

Model	Training (RMSE or accuracy)	Test (RMSE or accuracy)
Linear Regression	11,281	12,254
Ridge Regression	11,336	12,289
Lasso Regression	11,304	12,265
LDA	65%	65%
QDA	62%	55%
Random Forests	100%	66%
SVM	64%	64%
Logistic Reg	64%	54%
Sample Size	25,637	12,819

Discussion

After applying various methods, we found that the lowest RMSE we could reach was ~12,300 shares and the highest accuracy we could reach was 66%. Though better than random, this accuracy is not extremely high. The accuracy remains around this value, even when classifying with different threshold values.

Part of the challenge of this task is that the features are mostly about the *structure* of the article, rather than its *contents*.

In future work on this project, we would use the words in the articles as data. Our initial attempts on this have yielded promising results.

*https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity