Predicting and Evaluating the Popularity of Online News

{ He Ren and Quan Yang } Electrical Engineering, Stanford University

Background

Reading and sharing online news has become an important part of people's entertainment lives. Therefore it would be greatly helpful if we could accurately predict the popularity of news prior to its publication for social media workers (authors, advertisers, etc.). Our goal is to predict the popularity of a news post (measured by number of shares) based on various features (see Table I.). In this project, we attempted to apply linear regression, logistic regression, decision tree, SVMs, kNNs, KPLS and SVR (with different parameters tested) algorithms to make predictions (to classify the case as "popular" or "unpopular"). We also used PCA, forward/backward selection and Fisher correlation scores to select features and we compared our results in detail.

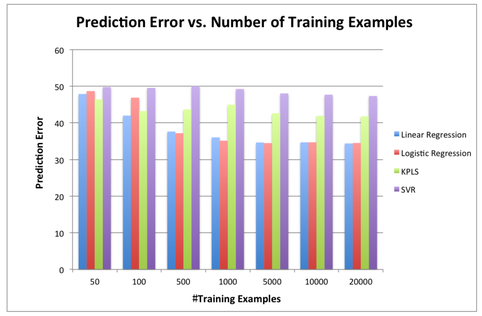
Data and Features

We used the dataset from UCI machine learning repository. It extracts 59 attributes describing different aspects of the article, from a total of 39000 articles published in the last two years from Mashable website.

TABLE I. All Available Features

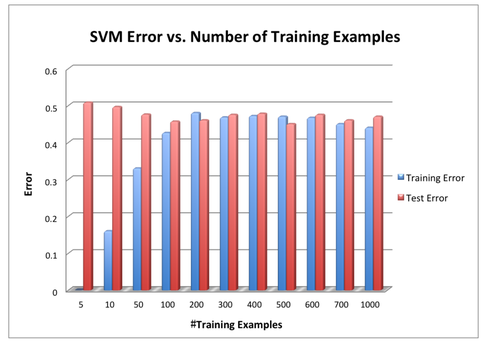
Aspects	Features
Words	Number of words of the title/content; Average word length; Rate of unique/non-stop words of contents
Links	Number of links; Number of links to other articles in Mashable
Digital Media	Number of images/videos
Publication Time	Day of the week/weekend
Keywords	Number of keywords; Worst/best/average keywords (#shares); Article category
NLP	Closeness to five LDA topics; Title polarity/subjectivity; Text polarity/subjectivity; Text sentiment polarity; Rate of positive/negative words; Polarity of positive/negative words; Absolute subjectivity/polarity level
Target	Number of shares at Mashable

Models

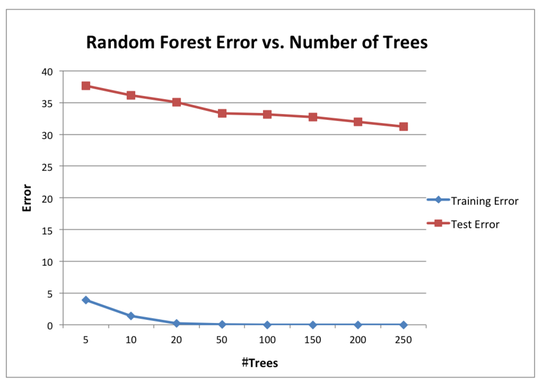


We tried over seven different algorithms on our data set, and observed their performance with respect to model complexity, number of features, and number of training examples.
The figure on the left shows the performances of linear regression, logistic regression, kernel partial least squares, and support vector regression. We can conclude that regressions, which is seemingly simple to implement, have better performances in fact.

The figure on the right shows the SVM test error and training error with the change of the number of training examples. We have used SVM with different kernel on all the 59 features. As we train the linear SVM model with more examples and apply k-fold cross validation, the training error and test error converges to a value slightly below 0.5. The linear model is considered to have high bias problem. Thus we try to use more complex kernels, which mitigates but still can't solve the problem.



Random Forest



In bagging (Bootstrap Aggregation), numerous replicates of the original dataset are created to reduce the variance in prediction. Random Forest use multiple decision trees built on separate sets of examples drawn from the dataset. In each tree, we can use a subset of all the features we have. By using more decision trees and averaging the result, the variance of the model can be greatly lowered. The figure indicates that the algorithm has high variance problem, so we can either increase the number of training examples and trees or use a smaller set of features.

Feature Selection

PCA:
We tried the PCA algorithm to reduce features dimension. However, the original feature set is properly designed and the noise (separate but correlated features) is limited. PCA doesn't give desirable result.

Filter Feature Selection:
1. Mutual information: We calculate the mutual information $MI(x_i, y)$ between features and class labels to be the score to rank features, which can be expressed as the Kullback-Leibler(KL) divergence:

$$MI(x_i, y) = KL(p(x_i, y)||p(x_i)p(y))$$

2. Fisher score: Fisher criterion is another effective way in feature ranking. The Fish score for jth feature is given by:

$$F(j) = \frac{(\bar{x_j}^1 - \bar{x_j}^2)^2}{(\bar{s_j}^1)^2 + (\bar{x_j}^2)^2}$$

- # 27. Avg. keyword (avg. shares)
- # 41. Closeness to LDA topic 2
- # 18. Is data channel 'World'?
- # 38. Was the article published on the weekend?
- # 16. Is data channel 'Social Media'?
- # 36. Was the article published on a Saturday?
- # 43. Closeness to LDA topic 4
- # 14. Is data channel 'Entertainment'?
- # 17. Is data channel 'Tech'?
- # 26. Avg. keyword (max. shares)

Main References

[1] Hensinger, Elena, Ilias Flaounas et al. "Modelling and predicting news popularity." Pattern Analysis and Applications 16.4 (2013): 623-635.
[2] "Predicting the Popularity of Social News Posts." 2013 cs229 projects. Joe Maguire Scott Michelson.

Contact Information

He Ren: heren@stanford.edu
Quan Yang: quanyang@stanford.edu