Autoranking Amazon Reviews

Ivan Gozali, Ilker Karakasoglu
Stanford University
{igozali, ilker}@stanford.edu

Overview

- Modeled the problem of predicting helpfulness scores on Amazon product reviews with the purpose of ranking them, as a regression problem.
- Evaluated the performance using the the coefficient of determination and rank correlation, and performed cross validation and regularization on several models.

Dataset

- Approximately 2.2 million reviews for Toys and Games products in Amazon in JSON format, containing information on reviewer name, product ID, helpfulness scores, reviewer's rating on product, title and review time.
- Filtered reviews using the following criteria: (1) reviews for products that have more than 10 reviews, and (2) reviews with more than 15 total votes, resulting in around 60,000 reviews.
- Using 80% of data set as the training set and 20% of data as the test set.

Feature Extraction

Outcome Variable: $\dfrac{\text{upvotes}}{\text{upvotes+downvotes}}$

Textual Features & Metadata:

- Review length, word count, unique word count, character count, punctuation count, sentence count, readability index, star rating

Bag of Words:

- Given a vocabulary of words V, if the j-th word in the vocabulary occurs c times in i-th review, then $x_j^{(i)} = c$.
- Top 3000 words are chosen and stop words are removed.

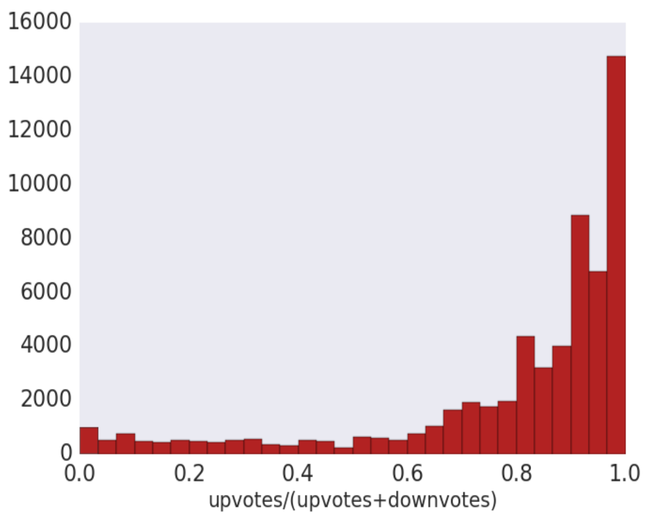


Figure: Histogram of review counts based on outcome variable

Results

Regressor	R^2_{train}	R^2_{test}	R^2_{CV}
Linear	0.34	0.26	0.19
Ridge	0.33	0.28	0.25
SVR	0.70	0.21	0.17
Random Forest	0.65	0.35	0.34

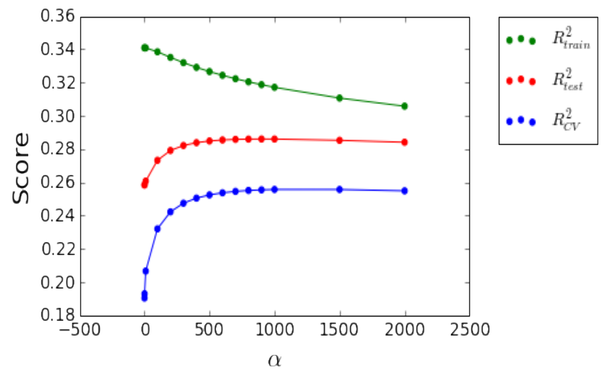
Table: R^2 values for train, test and cross validation on train data.

Linear Regression

- Although linearity assumption about the data may not be realistic, it still performs decently.
- Suffers from high bias.

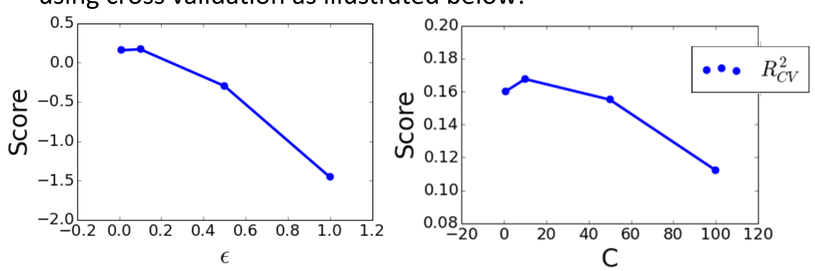
Ridge Regression

- To combat variance, make use of the penalty norm in ridge regression.
- We train a ridge regression model using various values of the penalization constant α.
- Trained on various α values to determine best fit, and obtained $\alpha = 400$.



Support Vector Machines for Regression

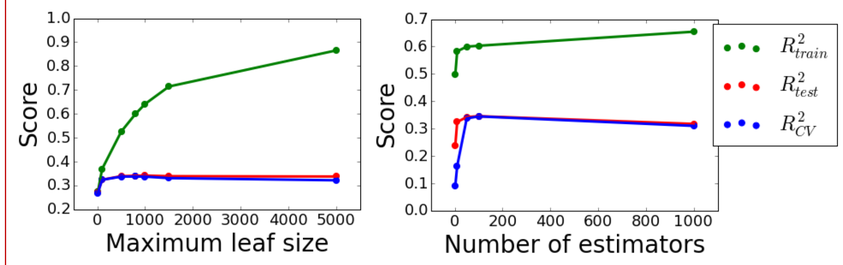
- A method effective in high dimensional spaces.
- Possible causes low performance: incorrect choice of kernel and skewness of data.
- Performed an exhaustive search over tunable parameters ε and C using cross validation as illustrated below:



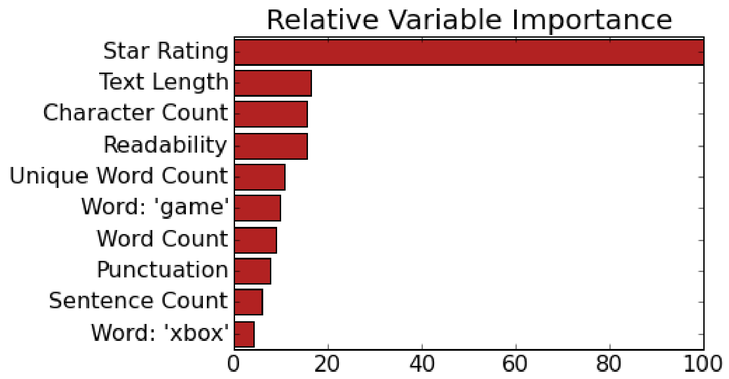
Results

Random Forests

- An ensemble method utilizing trees.
- Effective for decorrelating features and trees.
- A high-variance, low-bias method.
- Regular cross validation isn't required. Out-of-bag (OOB) error performs as cross-validation.
- Using OOB, performed a search over tunable parameters the maximum leaf size and the number of estimators as illustrated below:



- Did a variable significance test to evaluate the importance of used features. Textual features dominate the overall importance. Words relevant to the topic of the data set are also significant.



Conclusion

- The surprisingly high performance of linear regression and SVR with a linear kernel seems to indicate some linearity in data. However, random forests, a nonparametric technique, still performs better.
- The best regressor for the task is random forests with an $R^2_{train} = 0.64$, and $R^2_{test} = 0.34$.

Future Work

- Use more words and use PCA to perform dimensionality reduction.
- Tackle the skewness of data.
- Instead of using local representations (n-grams, bag of words), use continuous representations that capture semantic meaning better, such as the continuous bag-of-words model.