Applied Machine Learning
COMP 551

Luis Pinto
Rebecca Salganik
Mahyar Bayran

Assignment 1 Report

Prof. William L. Hamilton
McGill Faculty of Science

Winter 2019

Abstract

In this project we investigated the performance of linear regression models for prediction comment popularity on Reddit. We present different features to characterize the comments and analyze the accuracy of such proposed models. The best performing model contained the three given features, as well as set of unigrams containing the top seven words, a set of bigrams containing the top 24 words pairs, children count cubed and comment length. Our highest performing model had a test error of 1.2770407.

1. Introduction

A lot of work has been done to research the success rates of different models in predicting popularity on social media. For example, "*incorporating popularity in topic models for social network analysis"* (by Y. Cha, B. Bi, C.-C. Hsieh, and J. Cho, 2013) or "*Predicting the Popularity of Web 2.0 Items Based on User Comments"* (by X. He, M. Gao, M.Kan, and Y.Liu, 2014). However, most of the earlier papers focused on *YouTube* and other more prevalent social media platforms. Given a dataset containing 12,000 data points of Reddit user comments, we attempted to use similar methods to predict future comment popularity for Reddit.

2. Dataset

The dataset contained approximately 12,000 points. We chose to split it in the following proportion 10000:1000:1000 for the training, validation, and test sets (respectively). Within the dataset we were provided with three initial features: one which specified whether each comment was a root, its count of children, and controversiality ranking. In addition, we extracted a set of our own text features by initially creating a bank of all the words used in our dataset. Following this, we created a dictionary where each key/value pair contained a word and its use frequency. This allowed us to easily compute the top N words which were then converted into a matrix with a row for each data point and column for each of the N words. We then used this data to create another text feature which used a tool frequently used in natural language processing, bigrams.

2.1. Extra Features:

As a first extra feature we used bigrams to calculate popularity of word pairs (and thus predict the popularity using a combination of two words). N-grams are popular choices as features for Natural Language Processing projects. Using the *nltk* library we were able to pick the top ***B*** bigrams. For each data point, we calculate the frequency of each bigram. Combining this with the top ***C*** unigrams allowed us to make a text feature which greatly improved our predictions. Other extra features we added were the comment length and added a column in which we cubed each children count.

2.2. Ethics in Data Usage:

Working with data gathered from social media (without the explicit permission of the users) has many ethical ramifications. Due to the fact that posts are neither anonymous nor authorized, a case can be made that, in using this data, we are impinging on the Reddit users' rights to privacy.

3. Linear regression models

We used two methods to perform linear regression on the data: closed-form and gradient descent. The closed-form method is a least-squares fit approximation to find the solution to a system of linear equations. The gradient descent method is an iterative optimization algorithm that finds the solution by minimizing the error function. It requires as inputs the features matrix, the target parameter, a matrix of initial weights (w_0) and three hyperparameters (ε_0, η_0 and β). In this project, the values of w_0 are randomly generated numbers between 0 and 1 and the value of $\varepsilon_0$is kept at 0.001 to minimize the runtime. Moreover, we normalized the gradient descent with respect to the number of data points by introducing a factor of inverse length in the formula. Furthermore, in order to assess the efficiency between models, we use the mean-squared error.

4. Results and Discussion

4.1. Closed-form using the 3 simple (non-text) features:

The running time was found to be 0.00124s on average. This runtime includes predicting the weights and calculating the mean-squared errors on the training and validation sets. The reported mean-squared errors (MSE) for the training and validation sets are 1.0846830709157251 and 1.020326684843145, respectively. It is important to notice that the MSE using this method never changes.

4.2. Gradient descent approach using the 3 simple features:

The average running time was found to be 0.008s. This implies a significant time increase with respect to the closed-form solution as shown in *Figure-1*.

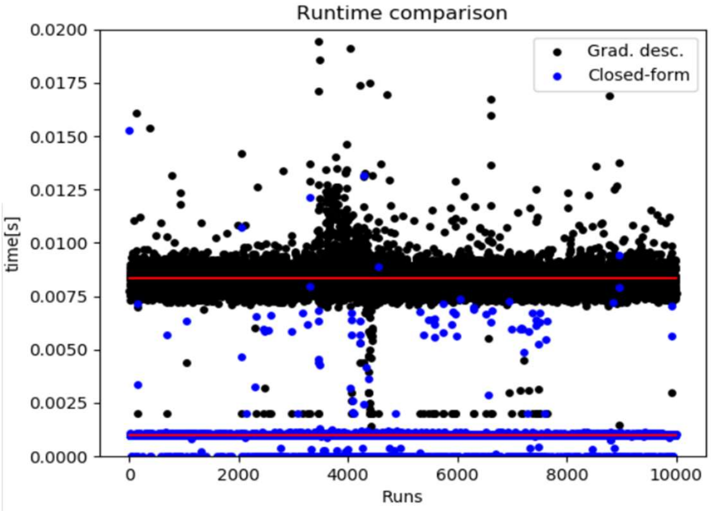

Figure-1. *Runtime comparison closed-form and gradient descent method in 10000 runs.*

Keeping the speed of the decay remains constant, various initial learning rates are tested on the training and validation sets to compare their performance. The data in *Figure-2* indicates that the minimum errors on both sets are found when the hyperparameter η_0 is in between 0.47 and 0.48.

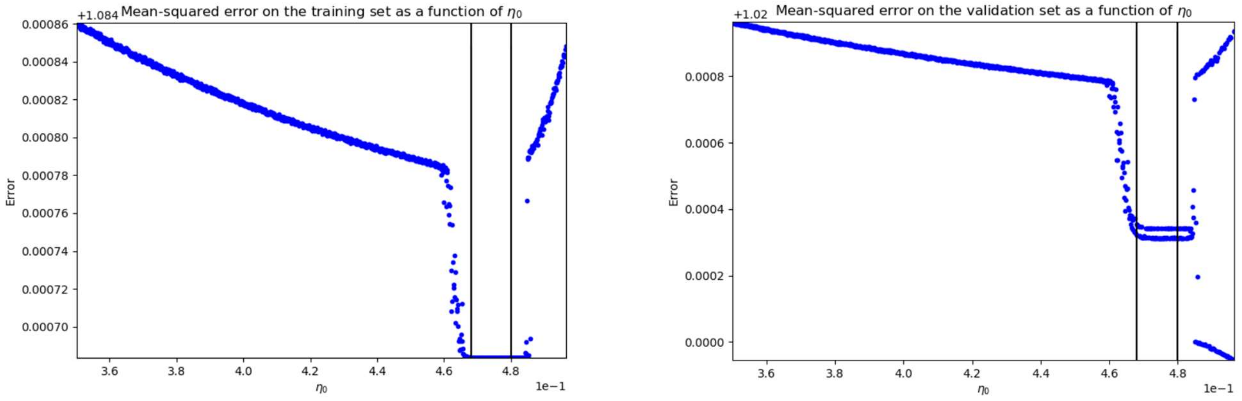

Figure-2. *MSE on the training and validation sets as a function of the hyperparameter* η_0

Moreover, different values for the decay rate in β are tested to minimize the MSE on both data sets. The next figures show a threshold of 0.3×10^{-6} imposed on the speed of the decay to keep the errors from changing significantly.

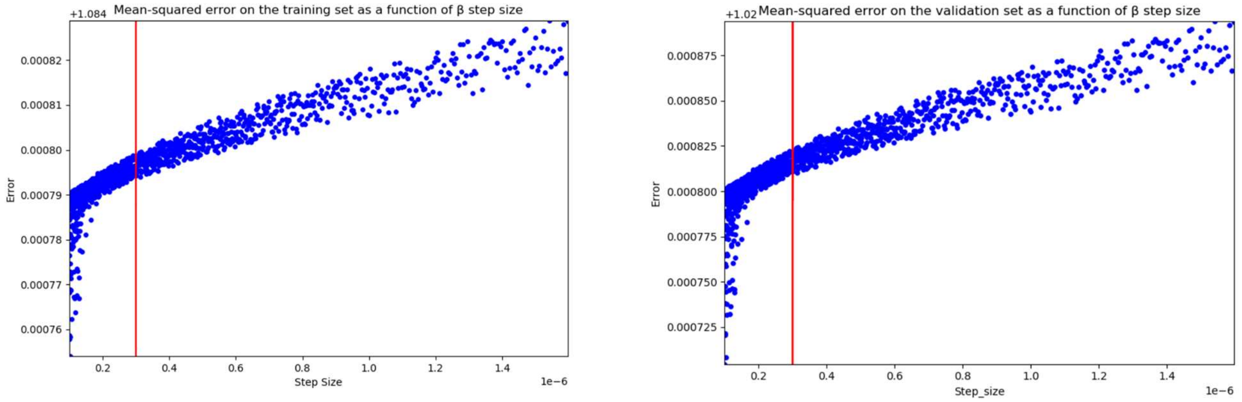

Figure-3. *MSE on training and validation sets for different values of the decay rate in* β

Considering the optimization on the hyperparameters mentioned above, the MSE for the training and validation sets are 1.0846830709157251 and 1.020326684843145, respectively.

4.3. Closed-form using the 3 simple features + extra features:

As explained above, the top 60 words and top 160 words were included in the features matrix to improve the performance of our model. This was implemented on the closed-form approach and the results are summarized in Table 1. It is difficult to say with certainty that absolutely no overfitting occurs. However, because our validation error is lower for all the tests (excluding the version which includes a feature containing the top 160 words) we can comfortably say that overfitting does not occur there. In the version with the 160 word feature we believe that there is overfitting.

Also, we found that the combination of the 3 simple features, 7 top words, 24 bigrams and the length of the comments gave us a low validation set MSE but failed to improve the MSE on the test set. However, if we add another extra feature to it, mainly children cubed, we found that the test set MSE did better as indicated in Table 1. Despite the good results, we found that the bigrams feature takes an extraordinarily long time to run. Therefore, it would probably not be viable for a larger data set or a system with time constraints.

Linear regression method	Number of top words	Extra Features	Training set MSE	Validation set MSE	Test set MSE
Closed-form	0	None	1.084683	1.020326	-
Gradient descent	0	None	1.084825	1.020886	-
Closed-form	60	None	1.060429	1.055807	-
Closed-form	160	None	1.047776	1.068390	-
Closed-form	7	Bigrams-length	1.072428	1.005290	1.307247
Closed-form	7	Bigrams-length-children cubed	1.029228	0.984993	1.277047

Table-1. *All models used as a minimum the 3 simple features.*

5. Conclusion

In conclusion, we found that the best performing model contained the three initial features, as well as a unigram set containing the top seven words, a bigram set containing the top 24 word pairs, children count cubed and comment length. Our highest performing feature had a test error of 1.2770407. Also, we had originally thought that gradient descent would be faster than our closed form approach, but we found the opposite. Furthermore, the key point that we drew from our testing was the importance of choosing features. We experimented with different features that, before testing, seemed like they would greatly improve the performance of our models but when implemented they did not: filtering out punctuation, remove stop words, considering whether a comment had external links, and many more. Finding features which can legitimately improve performance would require a deeper understanding of the underlying principles beneath user's interaction with Reddit. Also, testing if multiplicative combinations of our current features improve the predictability of our model is another direction for future investigation.

6. Statement of Contributions

Names specified in order of greatest contribution in each task:

- **Task 1:** Rebecca Salganik and Luis Pinto
- **Task 2:** Mahyar Bayran and Luis Pinto
- **Task 3:** Luis Pinto, Mahyar Bayran, and Rebecca Salganik
- **Write Up:** Rebecca Salganik, Luis Pinto, Mahyar Bayran