COMP 551 Assignment 1

Huiyuan Yang 260570687
Yiran Mao 260850827
Jingyuan Wang 260860682

February 1, 2019

1 Abstract

In this project, we implemented and tested several linear regression models using both closed-form and gradient descent approach. The task is to use comments data from Reddit to predict popularity of different comments. We found that with this scale of dataset, the close-form solution has both shorter runtime and better performance than the gradient descent function. Then we tested with different text features and extracted new features to compare each model and found out that the 160 word occurrence, comment length as well as the interaction term of *comment length* and *is_root* property improves the model largely. Also we tuned hyperparameters of the gradient descent approach to look for the best model and finally we get the lowest MSE of 0.9801 over the validation set and 1.2885 over the test set.

2 Introduction

Reddit is an American social news aggregation, content rating, and new form of internet discussion board. It has become increasingly popular in the past year having over two billion pageview per month. In 2018, more than 1.2 billion comments has been posted with 153 million submissions from its user. The purpose of the project is to predict the future popularity of a comment using different feature scenarios and different machine learning models. We first selected 160 words from the word count in descendent order as text features. Further investigation included combinations of build-in features (*is_root, child, controversially*), text feature, individual comment word count, bi-gram and sentiment analysis. From the our investigation, the closed form model generates the smaller error than the gradient descent model with the same features. Mean square error(MSE) on validation set is used to evaluate the performance of the model. Runtime, mean square error (MSE) comparison between closed form and gradient descent are also performed on the same features set.

3 Dataset

In this project, we split the dataset containing 12000 comments into three partitions: 10000 comments for training set, 1000 comments for validation set and test set respectively. The raw data has 5 categories: *popularity_score, children, text, controversiality, is_root. is_root* is a binary variable. *children, controversiality* are two numerical variable. *popularity_score* is the target.

3.1 Data Preprocessing

To construct the word frequency table, we first normalize the the text data by applying lowercase transformation. The second step is to tokenize the comments and construct the word frequency table. By inspection, the empty space accounts the second most frequent token in the word frequency table thus it was removed to further clean the data. We then selected the top 160 words from the word frequency table to build the text feature and count the word frequency for each individual comment. This text feature is stored in dictionary data structure along with other build-in features.

3.2 New Features Testing

We have come up with 4 new features in total with two of them related to text length and the other two about text content. It is easy to get intuition that text length and comment popularity are of significant relevance thus we counted character numbers and took the logarithm of it for better performance instead of simple word count. Also we created an interaction term of the *is_root* binary value and *text length* as a new feature. Then we tried to make use of text content and counted the 208 most popular bigrams which has appeared over 100 times in the dataset as a further training to word occurrence. In addition, we have noticed that there are lots of sentimental words and even abusive words in those comments, so we used certain tool to calculate the sentiment marks of each comment as another input variable.

Since the data used in this project are text written by human being, it is likely to incorporate certain person's name and links. As a result, researcher analyzing the data might make public of personal privacy. Apart from this, the act of using the comments on the Internet without the consents of Reddit users is not considered completely ethical.

4 Results

4.1 Comparison of Two Approaches

In the first experiment, we aim at having a quick and better understanding of gradient descent and closed-form approaches towards linear regression, so here only the three simple features are used. When comparing the runtime and performance of different methods, we found that the runtime of closed-form solution is shorter, and its mean squared error(MSE) over predicting popularity of comments on the validation set is slightly lower(Table 1).

Table 1: Runtime and performance of two approaches

Approach	Closed-form solution	Gradient descent
Runtime	0.000	0.031
MSE on validation set	1.0194	1.0198

Gradient descent learning rate = 0.0001, initial weight = 0.1

Also, due to the dependency on hyperparameters(learning rate and initializations), gradient descent approach is less stable(Table 2 and Table 3) and we found out the relationship between learning rate and MSE as well as initial w with MSE(Figure 1 and Figure 2). It seems like the closed-form approach is better on a small dataset. However, if the dataset is huge and the model gets much more complex, this approach would become expensive in terms of runtime.

Figure 1: MSE - Learning rate

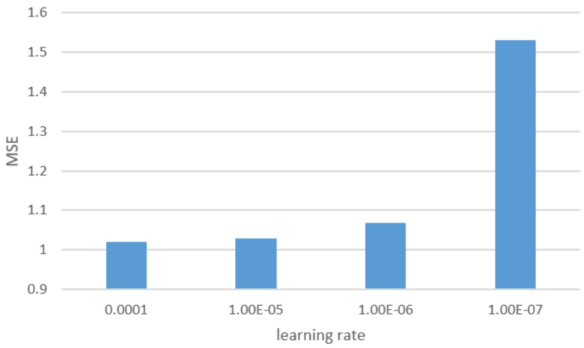

Figure 2: MSE - Initial w

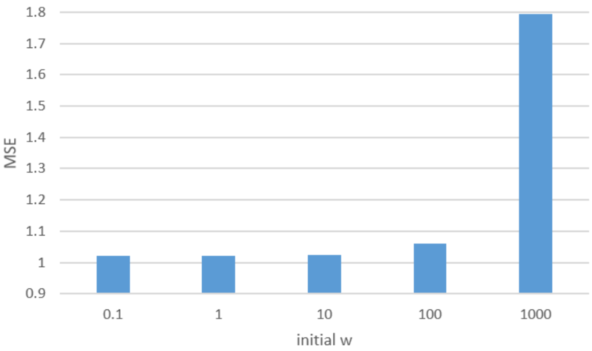

Table 2: Performance with different learning rates

learning rate	0.001	0.0001	1e-5	1e-6	1e-7
MSE (validation)	INF	1.0198	1.0289	1.0689	1.5294

Gradient descent initial weight = 0.1

Table 3: Performance with different initializations

Initial W	0.1	1	10	100	1000
MSE on validation set	1.0198	1.0201	1.0228	1.0595	1.7937

Gradient descent learning rate = 0.0001

4.2 Testing Text Features

Then we add text features to input variables to enhance our model and it is showing that the usage of text features has improved the performance of the model and reduced the MSE by 0.2 to 0.3. Using the full 160 word occurrence features and top 60 most common words have almost the same effect but still adding all 160 words features works the best and based on the results on training set and validation set, they are neither overfitting nor underfitting(Table 4).

Table 4: Text features

	No text features	Top-60	Top-160
Closed form MSE on validation set	1.0194	0.9826	0.9824
Closed form MSE on training set	1.0847	1.0612	1.0612

4.3 Usage of New Features

Finally, we incorporated our 4 new features: *comment length, the interaction of comment length and is_root, bi-gram word occurrence and sentiment analysis* and tested them on the validation set. We found that the first three new features we added has improved the performance in varying degrees. The performance of log of comment length has exceeded others, making an improvement by 0.046 in MSE and the usage of interaction term also reduced the MSE by over 0.007. However, sentiment analysis even cancelled out our improvement to the model's performance(Table 5). So *comment length* and *interaction term* are chosen to be incorporated in the best model.

Table 5: Incorporating new features

Features	MSE
No new feature	1.0564
Comment length	1.0106
Comment length+Interaction term	1.0033
Comment length+Interaction term+Bigram	1.0030
Comment length+Interaction term+Sentiment Analysis	1.0037
Comment length+Interaction term+Bigram+Sentiment Analysis	1.0034

When we run our best model on train, validation as well as test set, it is found that the closed-form approach performs better than gradient descent approach on training and validation set, however is worse on the test set(Table 6). In fact both two methods have poor performances on the test set maybe because the data instances is not identically distributed and we ignored the shuffle step in the very beginning.

Table 6: Performance of the best model

	Closed-form solution	Gradient Descent
MSE on training set	1.0457	1.0864
MSE on validation set	0.9801	1.0033
MSE on test set	1.3135	1.2885

5 Discussion and Conclusion

In this mini project, we used two main linear regression approaches of machine learning(closed-form and gradient descent) to predict the popularity of comments on Reddit. During this experiment, we mastered the basic process of machine learning such as dataset preprocessing and partition, feature selection, training and test. We practised tuning the hyperparameters such as learning rate and initializations to gain a better result.

Among several features we chose, it is found that the number of children and the length of comment are relatively effective. Due to the simplistic of dataset and model, the closed-form approach achieved better performance with lower MSE, shorter time and better stability.

Possible suggestions for future improvement include larger datasets, adding more useful features, as well as using more complicated model. Further data cleaning including removing article words, weblink and non alphabetical characters can possibly improve the prediction performance.

6 Statement of Contribution

During the experiment, the division of labour is clear-cut, each one being charged with specific responsibilities. Huiyuan Yang was responsible for data pre-processing and text feature extraction. Yiran Mao implemented training and test process. Jingyuan Wang took charge of the four new features and the experiment results. And we completed the final report together.