

# Machine Learning 101

Group 11:

Bide Xu (260711367)

Mengjia Gu (260790140)

Robin Luo (260851506)

Course Number: COMP 551

Course Name: Applied Machine Learning

February 1, 2019

### Abstract

Online posts and forums have become increasingly popular in daily use and valuable in marketing support. Hence, prediction on comment popularity is explorable for more applications. This project was purposed to assess the performance of linear regression models for predicting comment popularity on Reddit. Given the pre-crawled dataset, we applied two approaches of linear regression models, closed-form solution and gradient descent, to prediction models with and without text features. Although we modified value of hyperparameters in gradient descent and tried to improve performance, the closed-form solution actually runs faster and presents more accurate and stable results.

In addition, an improved version of gradient descent algorithm is also implemented, by introducing a new parameter called “distance rate”, which helps to let the learning rate value keep a reasonable distance from the epsilon value, so that the learning rate would not decay too quick. Therefore, this “distance rate” could help the algorithm run a reasonable enough time in order to improve the accuracy of the result (evaluated in MSE). This “distance rate” cooperates well with the Robbins Monroe logic, so could assist to achieve a well balance between run-time and accuracy.

Furthermore, new text feature based on 60 or 160 most high frequency words in the training data is implemented, with significant improvement for the prediction of popularity score. Also, other new implemented text feature such as “average words per sentence” indeed reduced MSE (mean square errors) of models and hence improved performance. Another new numeric feature is also implemented, which is the “cubic of children field” – *children*<sup>3</sup>. Both of these new designed features perform well in the improvement of the accuracy w.r.t both training data and validation data.

Moreover, we also implemented punctuation removing and stop words filtering features, though they did not bring obvious performance improvement according to our experiment.

## Introduction

Discussion board is frequently used in academic communication, product review and other public message exchange. This variety of information, which serves as a data source of the predicting technique, contributes to commercial development in recommendation system design(Rohlin 2016). Reddit is a classic and popular discussion platform, connecting communities with various topics. This project is proposed to construct and compare linear regression models in purpose of predicting comment popularity on Reddit. We are given some prepared data of selected comments, which still required preprocessing before import into experiments. Three features (children, controversiality, is\_root) and one output variable are directly from the original dataset. Other input variables, such as text features and word counts, were extracted or derived from text content of comments.

Given the pre-crawled dataset, we preprocessed data, extracted features, and built models for both closed-form solution and gradient descent. Not as the theoretical expectation, our experiment results illustrated that, compared to gradient descent method, closed-form solution is better in both runtime and accuracy. Compared models differently implemented with text features, we noticed that involving text content of a comment into models obviously improved the accuracy of prediction on popularity. By adding new input variables like

“average number of words per sentence”, we deduced the significance of comment content in prediction of popularity.

## Dataset

The original dataset consists of 12,000 data points in format of dictionary. Each dictionary contains five keys with different types of value. Data in each type of information forms one or more variables in linear regression models. We translated some of them in format better support linear regression models, as described below:

- popularity\_score: numeric output variable, larger value means more popular comment
- children: non-negative integers indicate number of replies received
- controversiality: binary metric, 0 means not controversial and 1 means controversial comment
- is\_root: binary, True represents a direct comment instead of being a reply, and contrary for False
- text: text content of a comment, which is convert to lower case and split by space, to generate new features

Besides the popularity\_score representing how popular each comment is, other information are taken as input variables that may affect prediction results. Text features are drawn from a wordlist of most frequently used words in all given comments. This wordlist contains 160 words ranking by occurrences in text content of comments. We as well count total number of words of comment and average word per sentence in each text value in order to add them as new features improving prediction.

After preparing features, we divided dataset into three parts for different purposes: first 10000 for data training, next 1000 as a validation set, and last 1000 to test model. Considered the limitation of data source, our project may lose accuracy lacking data of more features like post topics and timeline. Failures in measuring popularity score could also lead to missing objectiveness and reliability of our data source. Moreover, user preference is always difficult to detect, especially that Reddit targets on public audience with various backgrounds.

## Results

All our experiment results are stored as .json file, which will be submitted with the source codes and other documents.

## Closed-Form Solution v.s. Gradient Descent

Simply with non-text features directly extracted from dataset, closed-form solution without extra hyperparameter stands out for its stability and smaller mean square errors of both training and validation sets. Besides input variables, three hyperparameters affected results by gradient descent. Lowering learning rate took longer time to reduce mean square errors, while decreasing run time ended experiment before reaching small mean square errors, as discussed in Figure 1. The new implemented hyperparameter “distance rate” was added to introduce some balance between speed and accuracy, although it could not outperform the ideal model with closed-form solution, it indeed provided a well balanced solution to solve the conflict between runtime and accuracy, bring better accuracy, almost  $10^{-10}$  in MSE regarding the ideal weight vector  $[-0.22627679][-1.08584747][0.37536403][0.82092517]$  obtained by the closed form method, compared to the normal version of Gradient Descent method, as detailed in Figure 2.

## Word Count Features

Implemented with the closed-form solution, models using more text features from the wordlist took longer time to present lower mean square errors on both training and validation sets. After applying the word filters to remove punctuations and stopping words, mean square errors even dropped to some lower value. Our experiment results support this assumption, the 60 words version could improve the MSE of validation data from 1.02 to around 0.98, while the improvement of 160 words version does not as stable as the 60 words version.

## New Features

The “average number of words per sentence in each comment” feature obviously reduced the MSE of linear regression models in predicting comment popularity, from 1.020 to around 1.010 for the validation data. This text feature is selected due to its performance is better than some other text features we have attempted, such as “total number of words in the comment”, “total number of sentence” and “average length per word”. Although the “total number of words in the comment” feature does bring some degree stable optimization, but its improvement is less than 0.005 in MSE, and therefore is selected in our final model.

The new numeric feature of “cubic of children” is also helpful to optimize the accuracy, the improvement is from 1.02 to 1.00 for the validation data. All these two new added features also work well for the training data.

## Performance for Combined features

When the word frequency, the “average number of words per sentence” and the “cubic of children” features are integrated together, our experiment shows that their improvement can be combined well, which improve the MSE of the validation data from 1.02 to around

## References

0.94. This overall improvement is also verified by the testing data, whose MSE can be reduced from 1.29 to 1.26.

## Discussion and Conclusion

This project indicated better results of linear regression models with closed-form solution than which with gradient descent, which against the theoretical suggestion of better performance by gradient descent. One possible explanation is that small size of dataset, requiring only few operations, cannot reflect the advantage of gradient descent in dealing with huge data and complex operations.

More and better filtered text features improve the prediction accuracy. Other features derived from text data, like comment length and word counts per sentence, better predict the comment popularity as well. Newly added features are also related to text content of comments. Therefore, text content significantly affects the popularity of a comment. Meanwhile, more text features take models longer time to run experiments.

More features in reality should be considered as important variables affecting how popular a comment on Reddit is. The idea of predicting comment popularity facilitates in technical development in commercial area, especially through improving recommendation systems.

## Contribution Statement

All members brainstormed for data preparation, feature extraction, and model construction.

- Bide Xu: main python coding
- Robin Luo: models improvement and graph plotting
- Mengjia Gu: results test and summarize write-up

All participated in whole progress and learnt through this project.

## References

- Domingos, Pedro. 2012. "A few useful things to know about machine learning." *Communications of the ACM* 55 (10): 78–87.
- n.d. 2005. The Conversation Starts Here.
- Rohlin, Tracy. 2016. "Popularity prediction of Reddit texts." *Master Theses*, vol. 4704.

## References

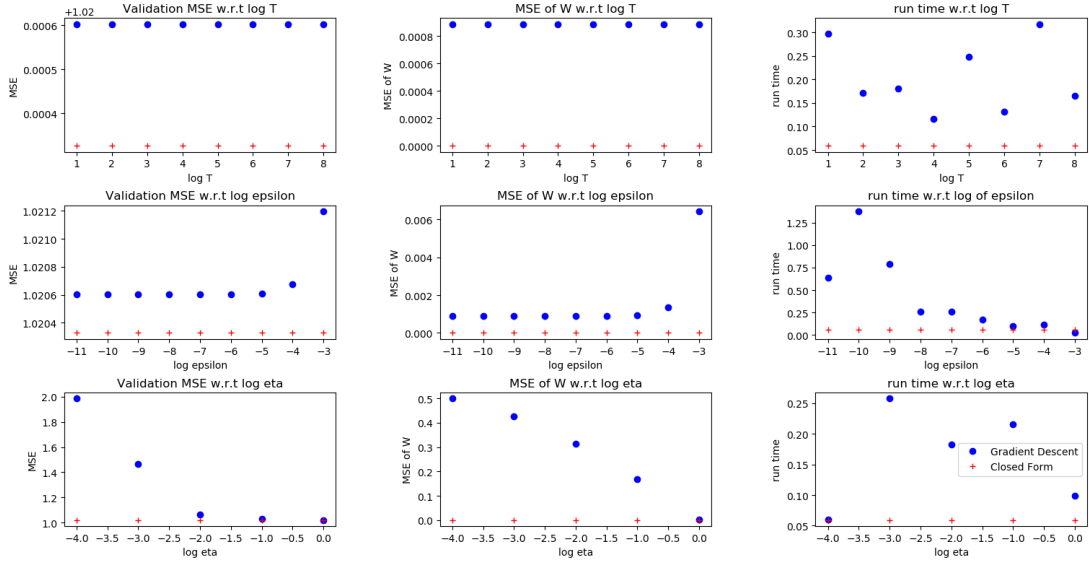


Figure 1: run time, accuracy performance comparison between closed Form Alg and Gradient Descent Alg, the stability of Gradient Descent Alg varies w.r.t different initial settings of its hyperparameters, such as period ( $T$ ) of Robbins Monroe,  $\epsilon$  and  $\eta$ . While the performance of the Closed Form Alg is stable.

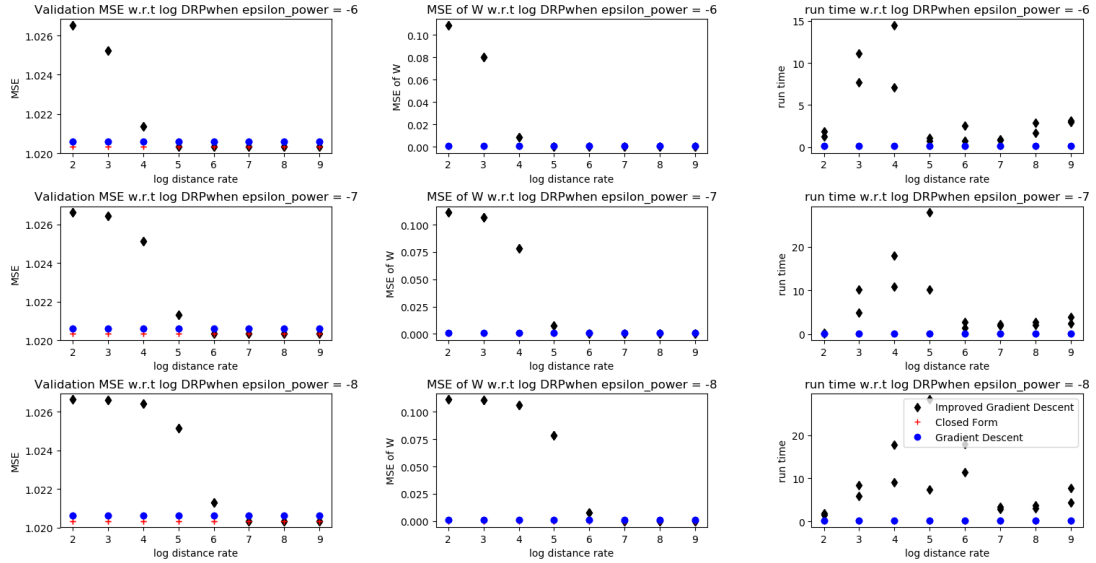


Figure 2: run time, accuracy performance comparison between closed Form Alg, Gradient Descent Alg and Improved Gradient Descent Alg, with  $\epsilon$  value around  $10^{-6}$  to  $10^{-8}$ . So the improved Gradient Descent Alg could be more accurate than the original Gradient Descent Alg, with acceptable more runtime