Reddit-Popularity-Predictor

MiniProject 1: Machine Learning 101

COMP 551 - Applied Machine Learning

Professor: Dr. William Hamilton

Authors: Group 37

Negin Ashouri, Le Nhat Hung, Sidi Yang

Abstract

Reddit (www.reddit.com) is a popular website where users can form interest-based communities, post content and participate in thread-based discussions. The objective of this project is to assess the performance of linear regression to predict the popularity of comments on Reddit. Gradient descent and closed form, which are two approaches of linear regression, are implemented on the data. Also, the performance of some primary features is analyzed. Besides these features, we proposed three new features of stemming, length of words and removing stop words feature representations. We observed that some of them improved the performance of the final model. We also found out that the gradient descent approach was slower than the closed-form approach.

Introduction

Reddit has recently become one of the most popular social platforms in the world. In this project, we will predict the popularity of comments from 12000 comments from **r/AskReddit**. There are some primary features provided, like "children" (which counts how many replies the comment received), "is_root" (which indicates whether the comment is a reply to another comment or not), "controversiality" (which describes how controversial the comment is) and the text itself. We added three other features which are comment length, stemming and removing stop words.

There are also some other papers that predict the popularity of comments on Reddit. However, they are working on a comparison between multi-class Naive Bayes, SVM and linear regression[1].

In our project, we use linear regression to implement prediction on Reddit comments. The approaches to performing linear regression are closed-form solution and gradient descent.

Dataset

We have a dataset contains 12,000 data points in total (the comments along with three features containing 1-controversiality 2- is_root 3- children). We partitioned the data into training, validation, and test splits. We used 10,000 points for the training set, 1,000 for validation, and 1,000 for the testing split. According to the first plot we took, which shows the features and their impact on the popularity score (Fig.1), and also the most frequent words' list, there are a lot of repeated stop words which can reduce our performance. There are also words with the same roots, which have the same meaning. Therefore, we decided to remove the stop words and also stem the text. Also, as well as taking a look at the most popular comment results, it seems that popular comments are usually longer. Therefore, we decided to add three new features to our data:

- **remove_stopwords [**i.e. "I" , "The" , "about" ,etc**]**
- **comment_length**
- **stemming:** [stemming is a process to reduce a word with prefixes and suffixes to its base form. For example, we can reduce "playing", "played", "plays" to "play" by implementing the stemming function. In our project, we used Porter Stemming function provided by Natural Language Toolkit Book (nltk) library. Thus, after stemming, we are re-define the most frequent words, and we will have 160 features without same words, which were previously considered as different words.]

One of the ethical concerns that may arise when working with public social media is that although people who have chosen to have public accounts expect their data to be public, it still does not necessarily mean that they agreed their data to be published, analyzed or judged. They expect some level of privacy[2]. More specifically in our case, although we are not using the username associated with each comment, all of these contents are publicly available on Reddit and may be misused by other people.

[1] Jordan Segall, Alex Zamoshchin. "Predicting Reddit Post Popularity." <http://cs229.stanford.edu/proj2012/ZamoshchinSegall-PredictingRedditPostPopularity.pdf>

[2] "But its already public, right?": The ethics of using online data | News & Analysis | Data Driven Journalism. (2019). Retrieved from http://datadrivenjournalism.net/news_and_analysis/but_its_already_public_right_the_ethics_of_using_online_data

Results

The following figure is a primary plot form the data containing three provided features along with the most frequent words and their impacts on the popularity score.

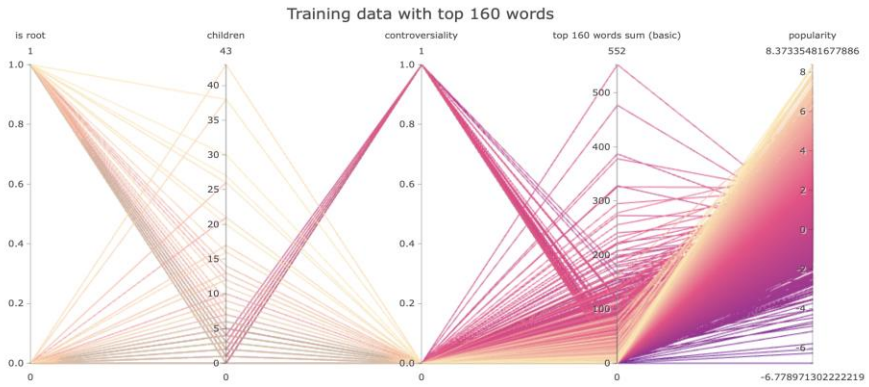

Figure 1. Training data with top 160 word

Following table is showing a comparison of the Runtime and MSE in gradient descent and closed form approach. (Only the first three features are considered)

Model	Hyper parameters (eps = le -6)	Run time (microseconds)	MSE	Iterations
Closed Form	beta = -4; eta_0= -5	423	1.0846830709157251	N/A
Gradient Descent	beta = -4; eta_0= -5	35793	1.3106408513933190	97963
Closed Form	beta = -5; eta_0= -5	456	1.0846830709157251	N/A
Gradient Descent	beta = -5; eta_0= -5	15635	1.1361887964459805	128869
Closed Form	beta = -4; eta_0= -6	521	1.0846830709157251	N/A
Gradient Descent	beta = -4; eta_0= -6	9681	2.0419981127161724	16407

Table 1. Closed Form and Gradient Descent Approaches for Different Hyper parameters

Figure 2 shows closed form and gradient descent prediction results with parameter beta = -4, eta_0 = -5 and eps = 1e-6 and the given popularity score.

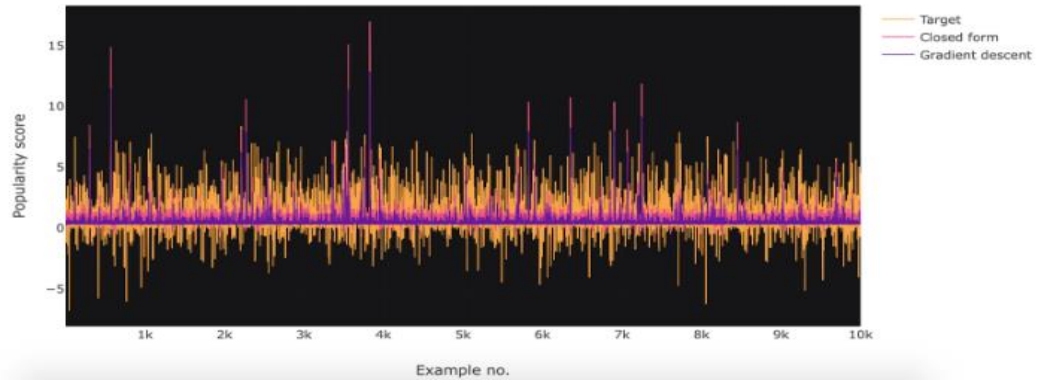

Figure 2. Popularity Prediction on Training Data

Observation and conclusion according to Table 1:

1- The closed form approach is more stable as it returns the same solution. Although, the gradient descent results are changing according to different values of hyperparameters.

2- In overall, the closed form model performed better than gradient descent (better MSE for both sets)

3- Closed-form has a better runtime during training (because of its straightforward computation), but there is still a risk of obtaining singular matrices

In the table below, we used the closed-form approach to compare a model with no text features, top 60 frequent words, and top 160 frequent words.

	Train Data	Validation Data
No text feature	1.0846830709157251	1.0203266848431447
60 words feature	1.060429141685383	1.0558077809463697
160 words feature	1.0477763217987115	1.0686390774956736

Table 2. The Accuracy of closed-form approach for different text features on training and validation

Observation and conclusion according to Table 2:
1-Judging from the MSEs, *top-60* is an incremental improvement over *no text*, and top-160 is a gradual improvement over top-60.
2- The fact that *top-160*'s training MSE is lower than *no text*'s demonstrate that *top-160* gives the best predictions out of the three.
3- Even though *top-160*'s validation MSE is the only one higher than its training's, its difference is not remarkable (1.0477763217987115 and 1.0686390774956736).
Therefore, in overall, adding the top 160 words features **improve** our model.

At the end after adding our three new features, we trained models and ran them on the validation set. Comment length feature improved our model (Table 3). However, we performed different algorithms of stemming (e.g., Snowball, Stem package from python and NLTK) to find the best model that improves our performance. The NLTK PorterStemmer function had the best performance.

Table 3 shows the MSE improvement of each of the new features individually.

Closed Form	Training Set	Validation Set
Before Adding New Features	1.0477763217987115	1.0686390774956736
Comment Length Only	1.0477454392805476	1.0673901584475689
Stemming Only	1.0466427040176542	1.0558077809463697
Remove Stopwords Only	1.042559330854879	1.0686390774956724

Table 3. MSE improvement of each feature

The following chart is demonstrating the performance of our best model on Validation and Test set.

	Top 160	Full(Top 160, length, stemming, stop words removed)
Training	1.0477763217987115	1.0434629526646237
Validation	1.068639077495673	1.1647577695495803
Test	1.3119720843275187	1.3607397144812439

Table 3. Closed Form MSE

Figures 4 and 5 show the prediction of popularity scores on validation and test data.

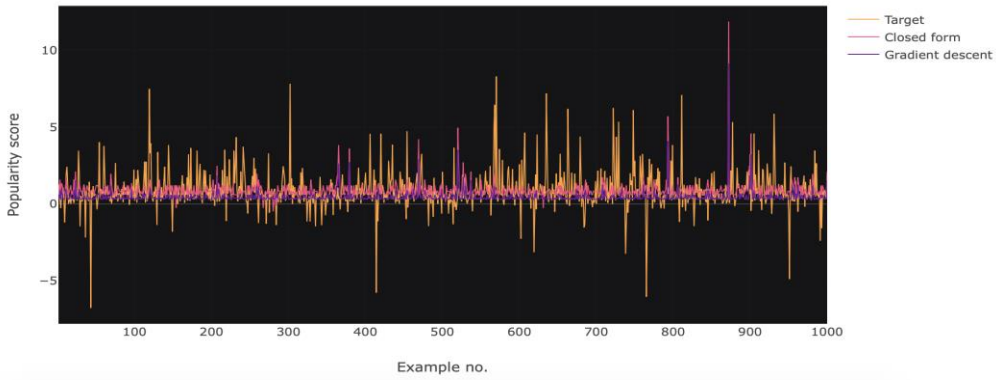

Figure 4. Closed Form Prediction on Validation Data

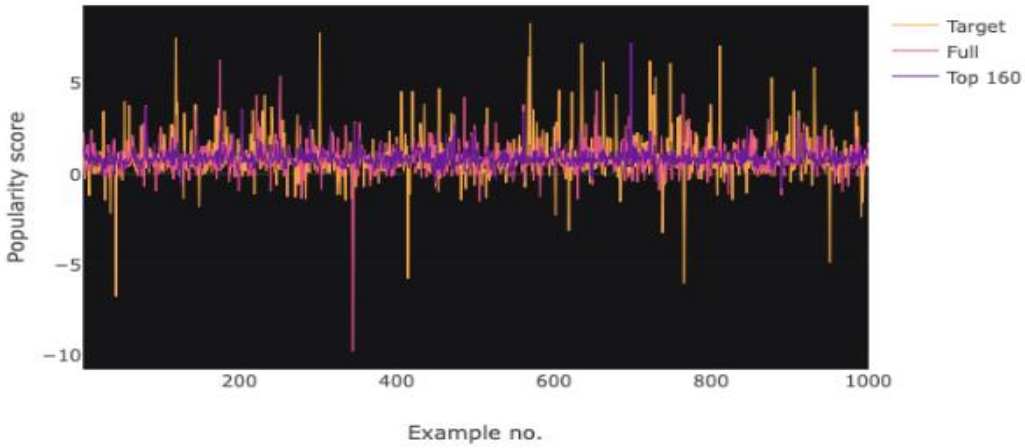

Figure 5. Closed Form Prediction on Test Data

Discussion and Conclusion

In our case study, the closed form approach is more stable and faster than gradient descent. Also, the160 most frequent words improved our model. Our three new features improved the model individually. However, our best model containing all the new features did not work properly on the test set.

After evaluating the results from the most popular comments, we observed that many of them have trending words. So we consider sending our data to google trend then use its results as our feature. However, as we are short in time and it takes lots of effort, it can be an option for future investigation.

Statement of Contributions

	Le Nhat Hung	Sidi Yang	Negin Ashouri
Brainstorming and Ideation	✓	✓	✓
Extracting the Data	✓		
Closed Form	✓		
Gradient Descent	✓		✓
Add 3 new features	✓	✓	✓
Experiment and Validating Added features	✓	✓	✓
Write- up		✓	✓