Machine Learning 101

Group 11: Bide Xu, Mengjia Gu, Robin Luo

McGill ID: 260711367, 260790140, 260851506

Course Number: COMP 551

Course Name: Applied Machine Learning

January 31, 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 preprocessed data, extracted features, and built models for both closed-form solution and gradient descent. Not as the theoretical expectation, out 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 comment text length and average number of words per sentence, we deduced the significance of comment content in prediction of popularity. Some more features like post topic, user community, and image use are considered as potential factors affecting comment popularity, which are not yet evaluated in this project.

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.

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. Including richer text features and considering comment length and average words per sentence indeed reduced mean square errors of models and hence improved performance. (With word filters, text features even help more in prediction of comment 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, out 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.

cite and figure

Results

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 increasing elapsed_time ended experiment before reaching small mean square errors. Although the third hyperparameter distance_power_rate was added to address some balance between speed and accuracy, the model using gradient descent still took longer time than the one with closed-form solution.

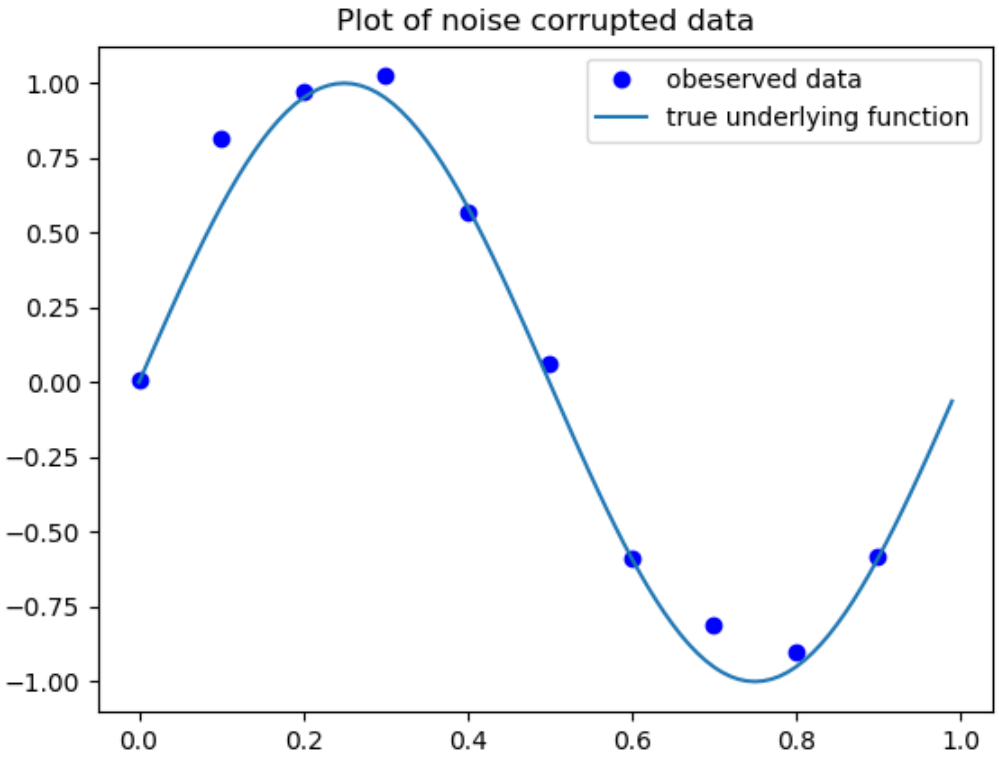

Figure 1: Relationship among

Text 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. Statistic data support this result in Figure 1.

New Features

Both total length of comment and average number of words per sentence in each comment reduced the mean square errors of linear regression models in predicting comment popularity. See Figure 2.

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 improves 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.

References

Domingos, Pedro. 2012. "A few useful things to know about machine learning." *Communications of the ACM* 55 (10): 78–87.

Rohlin, Tracy. 2016. "Popularity prediction of Reddit texts."

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.