

Machine Learning 101

Group 11: Bide Xu (260711367)

Mengjia Gu (260790140)

Robin Luo (260851506)

January 29, 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, our experiment results illustrated that closed-form solution is better in both runtime and accuracy than gradient descent method in runtime. 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**

Reddit is a classic and popular discussion platform, connecting communities with various topics. **reddit** 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, 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

7+ sentences + figures/tables

describe experiment results in task 3 + other interesting findings; must include:

1. runtime comparison between gradient descent and closed-form solution approaches
2. train + val mean-squared errors for a model without text features, a model with top 60 word text features, and a model with full 160 word occurrence features
3. train + val + test performance of a model that incorporates new features

Closed-Form Solution v.s. Gradient Descent

Text Features

New Features

Discussion and Conclusion

5+ sentences

summarize key takeaways from the project and possibly directions for future investigation

Contribution Statement

1-3 sentences