

Machine Learning 101

Group 11: Bide Xu, Mengjia Gu, Robin Luo

January 28, 2019

Abstract

100 -250 words

In this project we investigated the performance of linear regression models for predicting comment popularity on Reddt”, ”We found that XX features improved performance and that the gradient descent approach was faster/slower than the closed-form approach.

Introduction

5+ sentences

summarize project task, dataset, and most important findings (similar as abstract); extra background info and citations of relevant work (papers analyzing Reddit or predicting post/comment popularity)

Dataset

5+ sentences

briefly describe dataset(e.g. size, train/val/test proportions) and how to extract text features; describe new features come up with details; highlight any possible ethical concerns that might arise with a public social media dataset of this variety.

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

Discussion and Conclusion

5+ sentences

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

Contribution Statement

1-3 sentences