COMP 551 Project 1: Project Write-up

Yuxiang Ma, Pengnan Fan, Siyun Liao

Jan 31, 2019

Abstract

This project studies the performance and efficiency of different linear hypotheses on predicting the popularity of Reddit comments using criteria of least square loss and running time. To improve linear model, some extra experiments of features transformation, model selection and cross validation of parameters are also conducted. According to our investigation, we conclude that the closed-form approach have a shorter running time, higher stability, and lower Mean Square Error (MSE) than gradient descent, some of text features are redundant and possibly source of overfitting, and transformation of variables using exponential family functions (instead of log transform) and stepwise selection increased model performance on validation set.

Introduction

This project predicts comments popularities using linear models, and error was minimized using Method of Least Square. Original data was directly transformed into feature matrix X and label vector Y . Given a vector of features $X^T = (1, X_1, ..., X_m)$, linear regression outputs a prediction of Y using $\widehat{Y} = f(X) = XW$ (dimension of W: nx1). Best W in class of linear regression hypothesis is chosen using method of least square, which minimizes the least square error ($ERR(W) = (Y - XW)^T(Y - XW)$). Both gradient descent ($grad(ERR(W),\ W) =- 2X^T(Y - XW)$) and closed form ($\widehat{W} = ((X^TX)^{-1}X^TY)$) are used to compute an estimate of function f (denoted as $\widehat{f}$) that has lowest MSE on training data. Closed form shows a better stability, shorter runtime, and higher accuracy than gradient descent, and gradient descent have varied performance if hyper parameters were changed. In extra experiment, the best learning rate (β_0 , η_0 in $\alpha_k = \frac{\eta_0}{k+\beta_0+1}$) are 5-fold cross validated over training set, with varied (β_0 , η_0) in range. Models with 0/60/160 text features are also compared, result shows that 60 text features gives best performance on validation set. Children feature and text feature are transformed using a variation of exponential function to give a better result on validation set, when log transform is not able to be applied with input that contains 0 entries. Although fraction of nouns and whether comment contains URL prone to improve model, they did not show significant model improvement on validation set. Finally, best set of models among 160 text features (transformed and original), isroot, children (transformed and original), and controversiality were chosen using forward stepwise selection, which is a greedy algorithm that does constrained search ($m!$ times) on a tree of 2^m size of all possible combination of possible predictors. As a greedy algorithm, forward selection will not try all possible combination of predictors. This constraints picks up best predictors at each level, and returns the best (greedy) set, with a higher bias and lower variance as a trade off. (Hasite et al).

Dataset

In this project, a 12000-sample dataset is built and used to train our prediction model of the popularity of comments. This dataset is split into three parts: a 10000-sample training set, a 1000-sample validation set, and a 1000-sample test set. In each set, there are 3 basic features (isroot, children, and controversiality of a comment) and 160 word frequency features based on text. First, 160 most frequency words are learned from adding all comments in training set, and for each comment, count of each of the previous 160 most frequently words in comment (of each sample) became the 0-159 text features in that sample. For the transform of children feature using $g(X) = 1 - e^{-0.05X}$, this was came up with the scatter plot with children as horizontal axis and popularity as vertical axis, for shape of scatter fits the unit step response of first order LTI system (capacitor charging). This function is used as a substitute of log transform, for there are 0 in input, and log transform is not managed to deal with it without introducing bias. Hyperparameter of -0.05 (growth rate) was proven to improve the model the most on validation set. Similar as transform of all text features, according to scatter plots, $g(X) = e^{-0.04X}$ gives a good transform. Also, other set of features, including interactions among basic features and fraction of noun, verb, adjectives, were constructed, they are shown not to be significant in prediction.

In addition, for researches involving data directly from public social media, privacy should always be protected and treated carefully. The key to this problem is not to collect data with strong connection to someone. As researchers, we should not collect any personal identity data unless it is necessary to do so. If these data must be collected, then a level of abstraction should be applied, for instance, to collect only the last four digits of phone numbers.

Result

Comparison of Closed Form and Gradient Descent

Experiment 1: Comparison of Closed Form and Gradient Descent (Learn Rate: $\alpha_k = \frac{\eta_0}{k+\beta_0+1}$)

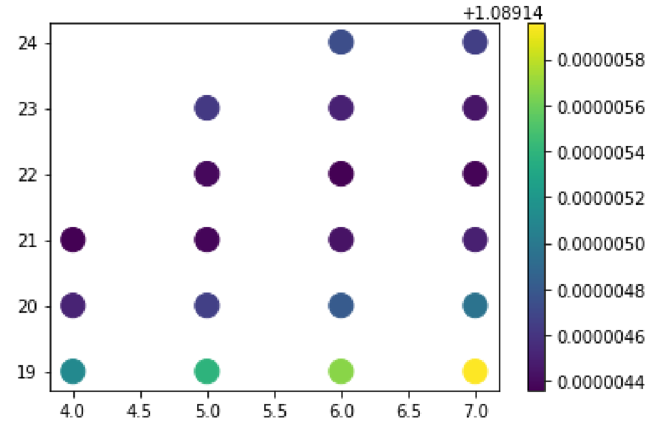
Result (Children, controversiality, is root, 1)

$X^T_{Closed\ Form} = (0.37536403,\ -1.08584747,\ -0.22627679,\ 0.82092517)$

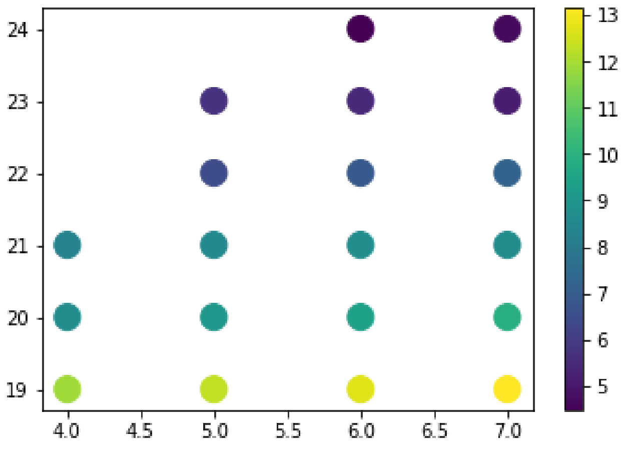
$X^T_{Gradient\ Descent} = (0.37532753,\ -1.07187479,\ -0.22618114,\ 0.82073393)$

MSE	Closed Form	Gradient Descent (0.0021, 4)
Training	1.0846830709157251	1.084685323915743
Validation	1.0203266848431447	1.0203862578853462
Running Time/s	0.29888367652893066	5.54496693611145

Extra Experiment: 5 fold Cross Validation of Hyperparameter of Gradient Descent ($\alpha_k = \frac{\eta_0}{k+\beta_0+1}$)



Cross Validation Score (MSE)



Averaged Runtime on 5 Training Sets

Note: horizontal axis should be interpreted as β_0 , vertical axis should be interpreted as $1000 * \eta_0$, and missing points indicates not possible to complete the computation.

Runtime: Gradient Descent has a significantly longer runtime than Closed Form.

Accuracy: Gradient Descent has a slightly lower accuracy on both sets than Closed Form, although it is not significant.

Stability: Cross Validation shows a varied MSE among different values of hyperparameter, which indicates that gradient descent is less stable than closed form.

Model Performance and Text Feature

Experiment 2: Performance of Linear Hypothesis in Different Word Frequencies

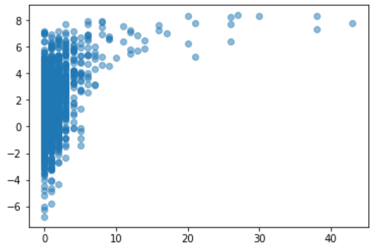
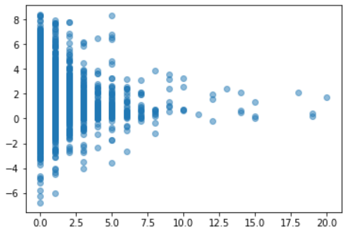
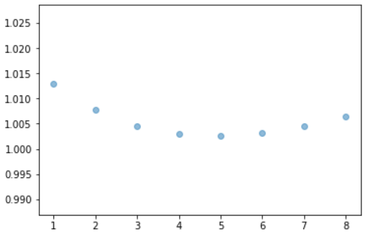
MSE	0	60	160
Training	1.084685323915743	1.060429141685383	1.0477763217987115
Validation	1.0203862578853462	0.9839397297217665	0.9950693970669264

On training set, as number of text features grows, error gradually decreases (-0.02), but validation error has lowest value at 60 features, and increased at 160 features. There is trend of overfitting on training set in 160 text feature, for training error decreased; however, validation error increased, compared with 60 features. There may be redundant features among last 100 features.

New Features and Final Performance

Experiment 3: Improvement of New Features on Hypothesis

To show improvement of new features, transform of children and text are added to model with 3 basic features and model with 3 basic and 60 text features, respectively.

		
Scatter Plot of Children and Popularity	Scatter Plot of Text and Popularity	Hyperparameter Choice and MSE on Validation Set

As the left graph shows, transformation analogs a capacitor charging function $g(X) = 1 - e^{-0.05X}$. As the middle graph shows, transformation analogs a capacitor discharge function (with stochastic noise A in $[-1, 1]$ as multiplier in $g(X) = Ae^{-0.04X}$). The hyperparameter of -0.05 was compared to be the best on validation set, according to minimum of graph on the right.

MSE	$g(X) = 1 - e^{-0.05X}$ of Children	$g(X) = Ae^{-0.04X}$ of 60 text features
Training	1.0420924933643285	1.058597084267706
Validation	1.0026119509501283	0.970215126374883
Improvement	0.01777430693	0.01372460334

Both of the new features actually improved more than 0.005 in terms of MSE, and does not show over fitting on training set.

Extra Experiment: Count Word Frequencies and Eliminate Stopwords

MSE	60	160
Training	1.0690435344142226	1.0504238189332942
Validation	1.0160834359569333	1.0377809673400562

Removing stopwords when counting frequencies does not improves model on validation set.

Extra Experiment: nounFraction, hasURL, and Word Count of comment

MSE	nounFraction	hasURL	Word Count
Training	1.0846431658913824	1.0846292895663272	1.083555207585464
Validation	1.0205294940039473	1.0206699968316335	1.0174970518644633

Although Word count slightly improved, the improvement is less than 0.005, and nounFraction and hasURL does not improve the model.

FINAL HYPOTHESIS and Extra Experiment: Forward selection

Forward selection was used to find the best set of predictors, and returned set of predictors consists of

Related to basic features

Transformed Children, controversiality, is root

Transformed Text Features

0, 1, 2, 3, 4, 5, 7, 9, 12, 13, 17, 18, 27, 32, 41, 48, 82, and 165th most frequent word.

Original Text Features

5, 9, 13, 18, 19, 20, 21, 23, 27, 28, 29, 30, 33, 36, 37, 38, 39, 41, 43, 44, 46, 51, 53, 54, 55, 58, 61, 64, 65, 68, 69, 70, 72, 75, 76, 79, 84, 85, 87, 90, 91, 92, 94, 95, 97, 100, 102, 103, 104, 110, 111, 113, 114, 119, 121, 124, 125, 126, 128, 129, 132, 139, 141, 142, 143, 152, 153, 156, and 158th most frequent word.

PERFORMANCE of FINAL MODEL

Performance	Training	Validate	Test
MSE	1.0154527650222982	0.9200926839324391	1.2736897834505778

This is an example of Meta-overfitting, since validation error is less than test error a lot, yet training error is higher than validation error.

Discussion and Conclusion

Gradient Descent has a longer running time and worse stability than closed form. Since the learning rate is small, number of iteration is higher than in closed form (given size of dataset), and matrix multiplication is also required in each iteration.

Redundant features have potential to overfit training data, and forward selection can effectively drop redundant features according to performance on validation set. In the 0/60/160 text features experiment, redundant text features fits training set so well that the model performed worse on validation set. Therefore, forward selection is used to greedily pick up combinations of good features with possibly lowest error on validation set. However, forward selection caused meta-overfitting on validation set, for testing error is larger than validation error in final model performance.

Exponential family functions is a good substitute of log function in transforming variables that contains 0. This transform actually improved model a lot. However, due to limited computation power, large-scale cross validation could not be performed on hyper parameter of exponential-family transforms on children and all text features.

In the future, runtime of gradient descent could possibly be improved by normalizing or scaling input data, and accuracy could be improved by using different initial vectors. In addition, regularization could be used in model selection to prevent meta-overfitting. Again, more cross validations should be done to select hyperparameter and prevent overfitting. In transformation of text features with $g(X) = e^{-0.04X}$, stochastic noise in range [-1, 1] could be used to fit data better. However, there is no valid mathematical validation of this approach.

Statement of Contribution

Yuxiang Ma: feature engineering, experiment design and excecution, code implementation, and report finalizing.
Pengnan Fan: report finalizing, report draft, suggestion on feature engineering and experiment design.
Siyun Liao: report draft.

Reference

Hastie et, al. (Second Edition) Chapter 3.3.2 Forward and Backward Stepwise Selection. *The Elements of Statistical Learning* (pp.58). Location: Springer