Assignment 1

Yung-Kuei Chen

yxc177030

- Data Prepare/Preprocessing

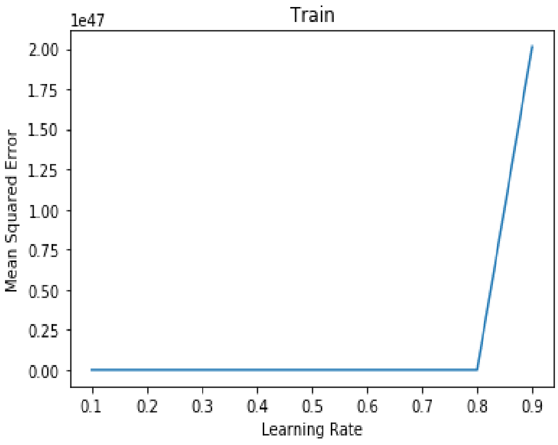
The reason why we need to divide the dataset into train and test sets sampling randomly is to create a generalized model. We have three following steps to achieve this. First, use pandas.read_csv to import the dataset. Second, normalize the data scale to decrease the discrepancy by applying MinMaxScaler in sklearn package. Third, split the dataset into training and testing set by 70% training and 30% testing via train_test_split function from sklean.model_selection. At the end, we have four datasets: X_train, X_test, y_train and y_test.

- Build linear regression model

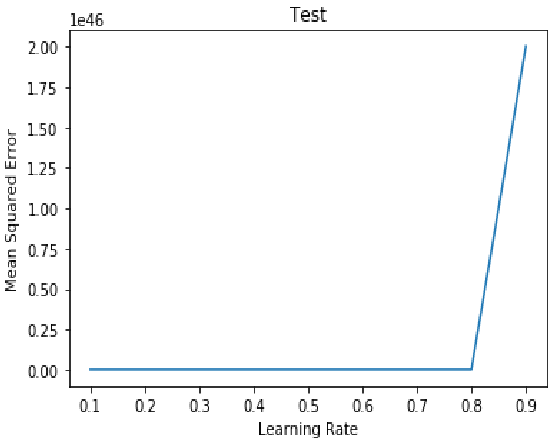
The concept of linear regression is to build a general model which can fit as more data points as possible. We randomly selected the initial coefficients to build the model, and then applying gradient descent to find the optimal solution for the model. The cost function we chose here is sum of squared errors. In every iteration in gradient descent, the goal is to find lesser and lesser cost than previous iteration.

- Linear regression model with different learning rate

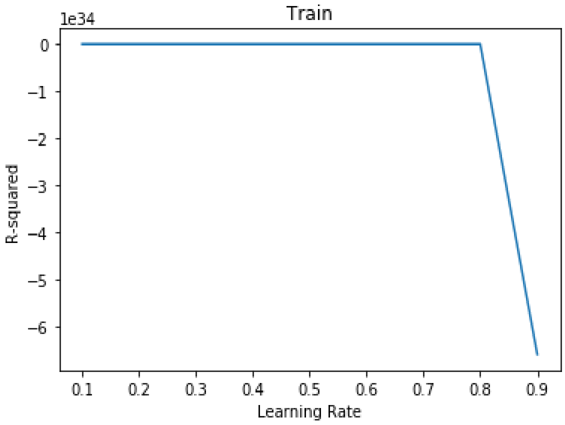
There are some parameters that we can assign to the model to have a better result. The learning rate is one of the important parameters we can adjust to reach the optimal point correctly. If the learning rate is too high, we will have overshooting problem in gradient descent. The mean squared error will be extremely high after some point. I ran a model using randomly 10 features with 9 different learning rates from 0.1 to 0.9 with 0.1 gap between now and the next rate. We noticed that at some point around learning rate is equal to 0.9 the MSE increases dramatically. That means this learning rate causes overshooting problem in this model. (Graph1&2)The R-squared value gets smaller dramatically which means the model does not fit the data. (Graph 3&4)



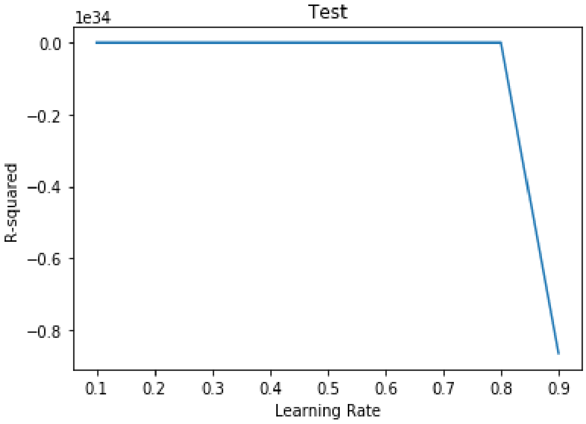
Graph 1



Graph 2

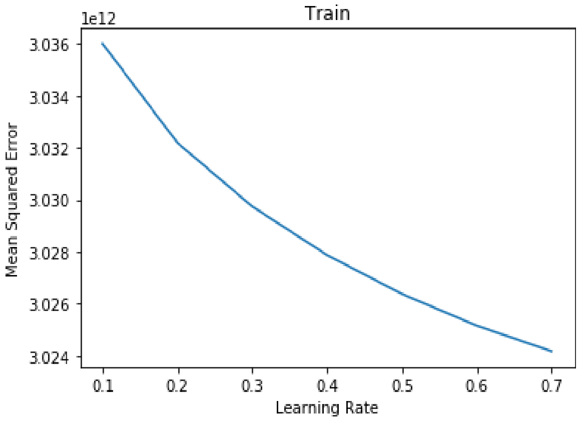


Graph 3

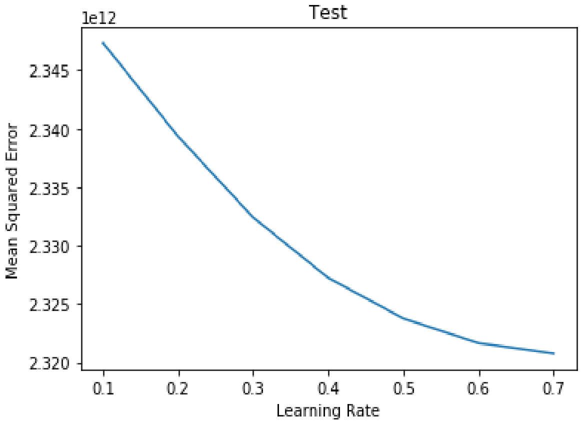


Graph 4

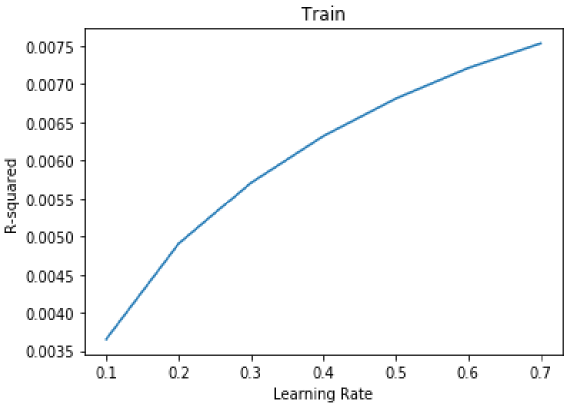
However, if we properly increase learning rate, we will get a better result in limit iteration. In graph 5 to 8, it represents the learning rate from 0.1 through 0.7 from previous model. For the higher learning rate in a proper interval, the MSE is getting lower which means the learning rate will lead model converge more quickly (graph 5 and 6), and R-squared value is getting higher which means the model fits data more (Graph 7 and 8).



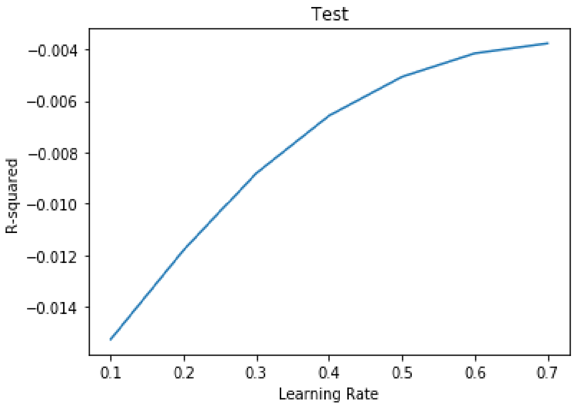
Graph 5



Graph 6



Graph 7



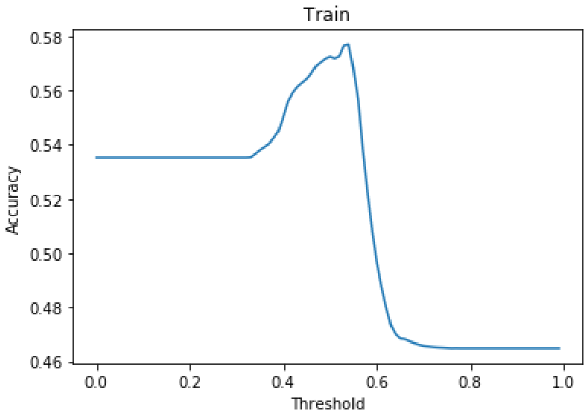
Graph 8

- Build logistics regression model

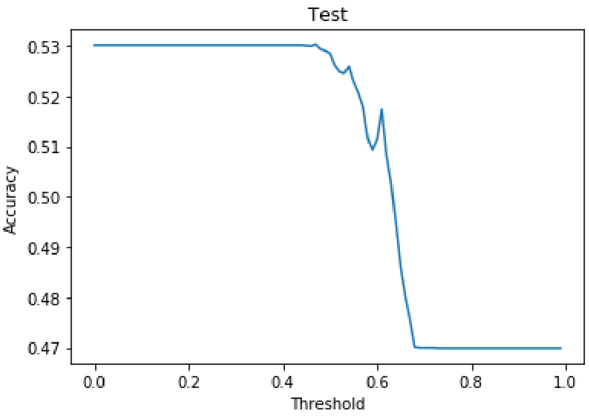
Same as build the linear regression model, using gradient descent to find the optimal solution for model. The difference from linear regression model is that the logistics regression uses different cost function: cross-entropy and needs a sigmoid function to transform the result from [-∞,∞] to interval [0,1] to have a better performance.

- Logistics regression model with different threshold

The threshold is an essential parameter in logistics regression which helps model to predict class. The threshold value in logistics regression is between 0 and 1. If the predicted value is under threshold, it will be classified as a group, else another. Thus, the threshold is a dominate parameter to decide which value belong to which class. I ran a model using the same randomly 10 features with 100 different thresholds from 0 to 0.99. The graph shows the accuracy with the respect of different threshold. The graph 9 from training shows that the best threshold is between 0.5 and 0.6 because the highest accuracy. However, the result from training data is not always suitable for testing data. In the graph 10, it shows that the best threshold for this dataset is under 0.4 which does not match the threshold from training data.



Graph 9



Graph 10

- Compare the model using all features with the model using 10 random features

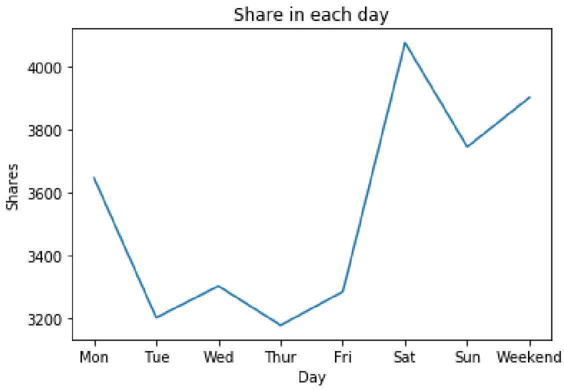
The features that I randomly selected based on the numpy random generator is: avg_positive_polarity, average_token_length, num_self_hrefs, max_positive_polarity, global_rate_negative_words, num_keywords, kw_max_min, n_unique_tokens, title_sentiment_polarity, LDA_00.

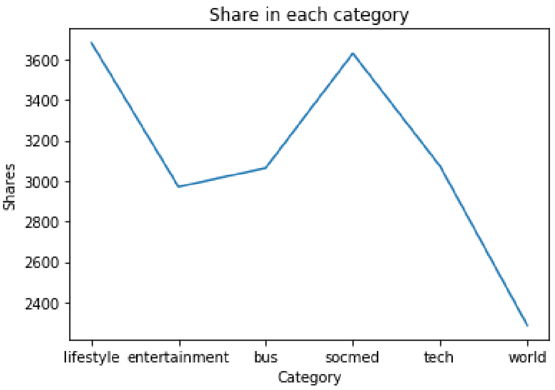
In the linear regression, the model with all features has a better performance than the model with 10 random variables. The MSE in from all features is lower and the R-squared value is higher than random. Moreover, in the logistics regression, the model with all variables also has a better performance with lower cost and higher accuracy. It is because that the model with all variable include more information.

Dataset	Model	Criteria	Random 10	ALL
Train	Linear	MSE	3.036e+12	**2.993e+12**
		R-squared	0.004	**0.018**
	Logistics	Cost	0.694	**0.672**
		Accuracy	0.535	**0.595**
Test	Linear	MSE	2.347e+12	**2.286e+12**
		R-squared	-0.015	**0.011**
	Logistics	Cost	0.717	**0.678**
		Accuracy	0.53	**0.588**

- Compare the model using 10 selected features with the model using all features and the model using 10 random features

  I selected the 10 features with following reasons:
  - Is_weekend: The articles posted on weekend may have higher shares because people have more free time to read posts on weekend.
  - Num_imgs: More pictures in article will explain more than the posts only with texts.
  - average_token_length: The short articles cannot explain any important idea. On the other hand, the long articles will scare the readers away. Therefore, it is better to control the articles in proper length
  - Num_videos: Videos tell more story than pictures and texts and are easier to understand which will make people put more emphasis on and share to others.
  - self_reference_avg_sharess: The referenced articles have higher shares means that people are interested in these kinds of topic. Thus, the shares of the original articles may have higher view and shares.
  - n_tokens_title: It is better to have proper length of title. If the title is too short or too long, there is no reader wants to read about it.
  - global_sentiment_polarity: The sentiment will impact how a reader feel after view the article. The positive sentiment in article will lead people to share the news.
  - global_rate_positive_words: People generally tend to share good news. If there are more positive words in the articles, more people will share these to others.
  - abs_title_subjectivity: If the titles are subjective, people will read the contents in these unbalanced titles to find out the reality.
  - data_channel_is_lifestyle: The lifestyle category has the highest share among all categories.





Dataset	Model	Criteria	Random 10	ALL	Top 10
Train	Linear	MSE	3.036e+12	2.993e+12	3.047e+12
		R-squared	0.004	0.018	-0.0001
	Logistics	Cost	0.694	0.672	0.676
		Accuracy	0.535	0.595	0.571
Test	Linear	MSE	2.347e+12	2.286e+12	2.311e+12
		R-squared	-0.015	0.011	0.0004
	Logistics	Cost	0.717	0.678	0.678
		Accuracy	0.53	0.588	0.57

The performance of the 10 selected features model is between two other models. In the linear regression part, the selected model is better than random model but worse than the model with all features. The reason why the model does not have great result is because that there is some information miss by only chose ten features. However, the selected model did a great job because it only includes 10 variables, and it is very close to the model with all variables. About the difference between random selected model and selected model is that the selected model has better performance on testing dataset which means that selected model is a more general model.

In logistics regression part, the selected model is better than random selected model but worse than the model includes all variables. The reason behind this is because that in the selected model we missed some important features which are essential to build an efficient model but still we selected some important variables which present not bad result when compare to the model with all features.