Comparison of Ridge Penalized Logistic Model and Ridge Penalized Simple Linear model in Dichotomous response

Sunghyun Cho

Seoul National University

sunghyun.sun.cho@gmail.com

January 10, 2017

Overview

1. Introduction
2. Methodology
3. Data Analysis
4. Conclusion

Overview

1. Introduction

2. Methodology

3. Data Analysis

4. Conclusion

Introduction

- logistic regression is one of the most widely used model for classification problem.
- Nonetheless, simple linear regression can also model dichotomous predictors using linear probability model.

- Goal : Comparison of Logistic regression and Simple linear regression in classification problem

Overview

1 Introduction

2 Methodology

3 Data Analysis

4 Conclusion

Methodology

- Linear regression with Ridge Penalization for Binary Data
- Logistic regression with Ridge Penalization for Binary Data
- Model Comparison Criteria : Area Under the ROC curve value
- Additional : GCV and GACV

Overview

1. Introduction
2. Methodology
3. Data Analysis
4. Conclusion

Data Analysis

- transaction information during May to July 2014 from a certain online trade intermediation site.
- In total, there are 3,802,601 transacions with 72,615 unique customers and 182 unique items.
- The goal is to estimate whether customers would purchase items on July based on the number of purchases during May and June.

- Training set : 50,000 unique customers with 2,536,915 transactions
- Test set : 22,615 unique customers with 1,265,686 transactions

Data Analysis

- Constructed the matrix of the number of purchases during May and June
- : the explanatory variable

- Constructed the matrix of whether purchased items on July for each set. If a customer purchased a item at least once, I coded as 1 and 0 if didn't
- : the response variable

- For the simplicity, instead of fitting multi-response model, fitted 182 single response models.

Data Analysis

- With the use of 5-fold Cross Validation method and AUC value, find the best λ
- Instead of using different λ for each model, decided to find the single optimal λ. For that purpose, calculated the mean of AUC values for each λ
- Ridge penalized Linear : best $\lambda = 0.2$
- Ridge penalized Logistic : best $\lambda = 4.2$

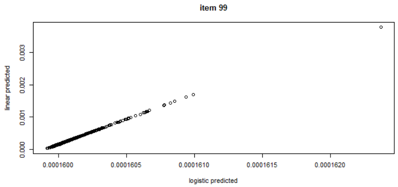
- With the use of GCV,
- Ridge penalized Linear : best $\lambda = 0.1$

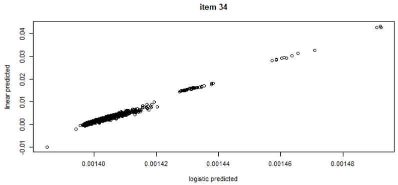
Data Analysis

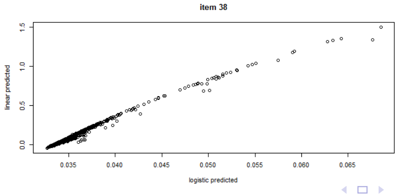
- Ridge penalized Linear : with the best $\lambda = 0.2$, test AUC $= 0.7635$
- Ridge penalized Logistic : with the best $\lambda = 4.2$, test AUC $= 0.7629$

- Shows quite similar result

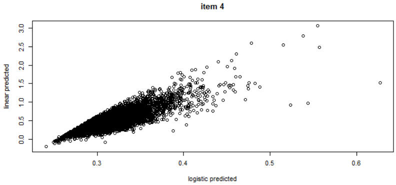
Data Analysis

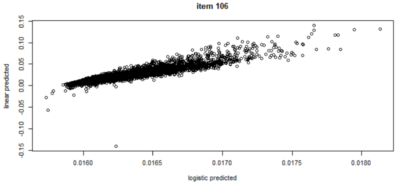


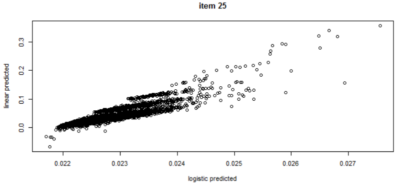




Data Analysis


item 4


item 106


item 25

Data Analysis

- The hit ratio of the top one item
- if we recommend the highest scored item for each customer, it is the proportion of who actually purchased each item.

- Ridge penalized Linear : 0.7344
- Ridge penalized Logistic : 0.6643
- Recommend the most frequent item : 0.6519

- two models clearly improved the result but the effect of logistic model was insignificant.

Data Analysis

- The itemwise ratio of top 20 percent hit ratio to overall hit ratio
- if we recommend the top 20 percent scored item for each customer, how many times the hit ratio is increased compared to when we recommend all items.

- the mean of the value
- Ridge penalized Linear : 2.8140
- Ridge penalized Logistic : 2.8097

- the linear model performs better than the logistic model in 85 items
- the logistic model performs better than the linear model in 45 items
- two model shows same ratio in 52 items.

Overview

1. Introduction
2. Methodology
3. Data Analysis
4. Conclusion

Conclusion

- Based on the mean AUC, two analysis produces quite similar results when applied to the same data set.
- The scatter plots of predicted values from linear and logistic regression also tell us two methods produces similar results in most of cases.
- In terms of recommendation problem, the linear model shows better results in general but two models are clearly better than suggesting the most frequent item.