SML: Exercise 2

Finn-Ole Honer, Thao Le, Jason Wang, Ramon Snellen

Contents

1 Introduction

This report aims to find the best set of predictors for past cumulative grocery sales (in dollars) for Dominick's Finer Foods.

2 Data

The data set contains seven years of store-level data collected at Dominick's Finer Foods by the University of Chicago Booth School of Business. The data can be found at https://www.chicagobooth.edu/research/kilts/datasets/dominicks. The data set contains 50 variables, which stem from:

1. customer count files, which contain information about in-store traffic;
2. a demographics file, which contains store-specific demographic data;
3. number identification files, which contain product information.

Of the fifty variables, `GROCERY_sum` is used as dependent variable. Furthermore, four categorical variables are dropped; `STORE`, `CITY`, `ZIP` and `SPHINDX`. The remaining variables are potential predictor variables.

3 Method

To find the optimal set of predictor variables, and there corresponding weights, we use a regression method that penalizes the size of coefficients. The penalty is useful when predictors are collinear, or the number of predictors destabilizes estimation. Let $P(\beta)$ denote a general penalty function. Then, the penalized regression equation becomes

$$L(\beta) = (\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) + \lambda P(\beta).$$

It follows that λ is the hyperparameter that determines the strength of the penalty. When $P(\beta) = \beta^2$, the regression is called 'ridge' regression. Similarly, when $P(\beta) = |\beta|$, the regression is called 'LASSO' regression. Finally, any combination $P(\beta) = \alpha|\beta| + (1-\alpha)\beta^2$ of 'ridge' and 'LASSO', where α denotes the weights, is called 'elastic net'.

4 Results

5 Conclusion and Discussion

6 Code

[REFERENCES]