

SML: Exercise 2

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

2022-11-05

Introduction

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

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.

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. This data set only consists of 77 observations for 50 variables, hence the number of predictors would destabilize estimation if not penalized.

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 convex combination $P(\beta) = \alpha|\beta| + (1 - \alpha)\beta^2$ of the 'ridge' and 'LASSO' penalty, where α denotes the weight on the 'LASSO' penalty, is called 'elastic net' (compare Zou and Hastie 2005).

Results

Conclusion and Discussion

Code

References

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20.