MATH 4995

Senior Project for CLA

Instructor: Wenqing Hu

An overview on linear model selection methods

Wenjing Yang

1. Introduction

Linear model selection methods are a critical part of statistical analysis with a wide range of applications, particularly in high-dimensional data analysis, where the number of variables is much larger than the sample size. The applications of linear model selection methods expand across various fields, from problems in economics and finance to problems in cancer research. In the former areas people consider optimizing portfolio performance by reducing the dimensions from hundreds of thousands of parameters to a small number that produce the best predictive model with small estimation errors (Fan, 2011). The latter areas cover problems in cancer research such as identifying the model that selects important genes microarray datasets to analyzing prognostic factors, therapeutic efficacy, recurrence and metastatic in breast cancer (Lonning, 2007; Huang, 2003). Approaches to variable selection methods with linear regression are evaluated based on its prediction accuracy and model interpretability. The least squares estimates are infeasible in fulfilling these two criterion in that they often result in large variance and do not

select a subset of variables that best represents the true model (Hastie, Tibshirani, & Friedman, 2009). In cases with constraints of limited accessible sample data, the challenges of selecting the set of explanatory variables that give the sparse representation to the true model arise. These explanatory variables must produce consistent predictive model. Adapting effective methods to selected variables that are of interest to a particular study is thus crucial.

Various methods and algorithms have been introduced in variable selection. This paper focuses on presenting an overview of Chapter 3 in the book, *The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009), which provides a brief introduction to some basic linear methods for regression. The remaining sections of the paper are organized as follows. We examine closely at the different types of model selection methods in Section 2. These methods are then classified as models that fall under the discrete process and models that satisfy the continuous process. In Section 3, we take a closer look at the discrete and continuous processes and compare the two procedures. Some concluding remarks are provided in Section 4.

2. Model selection methods

Consider a linear model with p predictors: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \epsilon$. There are various methods in selecting the model that gives the best prediction for the output of Y. The least squares method, a general approach often taken, is selecting the set of coefficients that minimizes the residual sum of squares (Hastie, etc., 2009). Nevertheless, when the number of

predictors outweighs the sample size, alternative approaches are often considered in order to control the variance. We take a closer look at the two different processes in model selection methods: the discrete process and the continuous process.

2.1 The discrete process

The discrete process discussed in this chapter by Hastie entails selecting a subset of predictors that generates a model with the least prediction error relative to the full model. The model selection methods in this case are discrete processes in a sense that the variables are either retained or removed. The least squares regression evaluates the coefficients of each variable and assess whether they are kept (Hastie, etc., 2009). Several model selection methods discussed in this chapter of the book are classified as the discrete process include best subset regression, forward-stepwise regression, and backward-stepwise regression.

The best subset regression method starts with a null model that contains zero predictors to start the prediction of the sample mean for each observation. Given a specified set of predictors, the best subset regression identifies the subset of the same size that results in the largest coefficient of determination (R-squared) or smallest residual sum of squares. The best-fitting model is obtained from a criterion (cross-validated prediction error, BIC, etc.) that chooses the model that minimizes the prediction error (Hastie, etc., 2009).

Instead of taking all possible subsets of predictors into account, which may be infeasible when the number of predictors is relatively large, the stepwise regression method adds or

removes each predictors from the whole set one at a time based on their statistical significance. There are forward stepwise selection and backward stepwise selection methods. The forward stepwise selection method uses a greedy algorithm that starts with a null model containing zero predictors. It then iteratively adds in predictors one at a time until a model with the smallest residual sum of squares or largest R-squared value is obtained (Hastie, etc., 2009). Reversely, the backward stepwise selection method begins with the full model containing all of the predictors, and removes each variable one at a time until the model with the smallest Z-score is obtained (Hastie, etc., 2009).

2.2 The continuous process

The continuous process differs from the discrete process in that the fitted model includes all of the variables but regularizes the estimated coefficients, with each variables shrunken towards 0 relative to the least squares estimates (Hastie, etc., 2009). Some shrinkage methods introduced in this chapter include the ridge regression and the least absolute shrinkage and selection operator (Lasso). These methods perform a proportional shrinkage and penalizes on the coefficients of each variables.

The ridge regression method includes a tuning parameter that controls the different impact for each variable with a shrinkage penalty that the estimates of coefficient estimators towards zero. According to Hastie (Hastie, etc., 2009), the ridge regression aims to minimize the function:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \; subject\ to \sum_{j=1}^{p}\beta_j^2 \le s$$

The best model from ridge regression includes all of the predictors, which may be a slight disadvantage in some cases. In comparison, Lasso is a good alternative to the issue. Lasso is a commonly used variable selection method proposed by Tibshirani (Tibshirani, 1996) that enhances the performance of predictive models by identifying significant predictors of a fitted model in linear regression problems. It is based on the technique of minimizing residual sum of squares to produce a number of coefficients approximately zero as the penalty parameter increases, and shrinking some coefficients to zero when the penalty parameter is suitably large (Zou, 2006). According to Hastie (Hastie, etc., 2009), the goal can be summarized as minimizing the function:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \; subject\ to \sum_{j=1}^{p}|\beta_j| \le s$$

Lasso also shrinks the estimated coefficients towards zero but the penalty term can lead to some estimated coefficient to be zero when the tuning parameter is amply large, and thus can select a subset of variables that gives the best representation of the true model.

3. Discussion & Conclusion

Generally speaking, linear models demonstrate a good predictive performance. Two key components when assessing a linear model is its prediction accuracy and model interpretability (Hastie, etc., 2009). Prediction accuracy is assessed based on the variance;

on the other hand, model interpretability is based on the number of relevant features selected from a certain method. Based on a simple linear model with ordinary least squares fitting method, various alternative fitting methods and procedures can improve the predictive performance, each having their own advantages and disadvantages.

In discrete processes the methods either include or exclude variables, which may be beneficial for interpretation when the goal is to select a subset of variables that best represent the true model. On the other hand, in the methods under the continuous process such as the ridge regression method, the tuning parameter plays an essential role in that it smoothly shrinks the estimated coefficients for each variable in a continuous range. This has a better effect in minimizing prediction error, relatively to the previous methods such as best subset regression selection method. The lasso method benefits intermediately in that it shares some properties under both the best subset regression method and ridge regression method. Overall, these fundamental model selection methods provide advancement in optimal modeling, which may lead to progress in real application across various disciplines.

Reference

Fan, J., Lv, J., Qi, L.: Spase High-Dimensional Models in Economic. Annual Review of Economics, 3(2011), 291-317.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. Springer.

Huang, E., Cheng, SH., Dressman, H., Pittman, J., Tsou, MH., Horng, CF., Bild, A., Iversen, ES., Liao, M., Chen, CM.: Gene Expression Predictors of Breast Cancer Outcomes. The Lancet, 361 (2003), 1590-1596.

Lonning, P.E.: Breast Cancer Progonostication and Prediction: Are We Making Progress? Annals of Oncology, 18 (Supplement 8) (2007), viii3-viii7.

Tibshirani, R.: Regression shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society, Ser. B (58) (1996), 267-288.

Zou, H.: The Adaptive LASSO and Its Oracle Properties. Journal of the American Statistical Association, 101 (2006), 476, 1418-1429.