Linear Model Selection and Regularization

Standford University - Statistical Learning Course

Seung Hyun Sung

11/24/2021

Objectives: Consider approaches for extending the the linear model framework. Improving the least squares by appropriate feature selection or shrinkage of feature coefficients to generalise the model in order to accommodate non-linear, but still additive, relatioships.

Contents

Despite its simplicity, the linear model has distinct advantages in terms of its interpretability and often shows good predictive performance. In this chapter we will uncover some useful alternative fitting procedures as a replacement to the ordinary least squares fitting.

Why even consider alternatives to OLS?

- **Prediction Accuracy**: especially when $\mathbf{p > n}$, to control the variance. Case where more features than the number of observations are present, one cannot use full least squares because solutions not even defined. Such condition could also raise over-fitting issue.
- **Model Interpretability**: By removing irrelevant features (insignificant of its corresponding coefficient estimates) more interpretable model is obtained.

6.1 Introduction to Feature Selection Methods

Subset Selection Identify all the possible subset of the p predictors that we believe to be be related to the response. Best subset selection, most productive set of features among all the possible combinations of features.

[1] M_0 -> null model {Intercept: simply predicts the sample mean for each observation}

[2] For k = 1,2,... p: fit all p^k models that contains exactly k predictors.

[3] Pick the best subset M_k among $M_0, M_1, ..., M_p$ defined as having smallest RSS or equivalently largest R^2.

The limitations on subset selection method

- **Computational reason** when working with large p.
  - p^k models an be problematic once k exceeds 20 number of features (below 10 is reccommended) as it computes every single possible combination of features.
- **Statistical problem**: the chance of finding a single best subset out of all the possible combination of features are likely to rise "over-fitting" issue.
  - Pays in price in variance {bias-variance trade-off?}

Suffer from statistical problems when p is large: larger the search space, higher the chance of finding models that look good on the training data, even though they may not have any predictive power on future data. Thus an enormous search space can lead to **overfitting** and higher variance of the coefficient estimates.

Stepwise The stepwise methods are discovered to overcome the best subset model where restricted to working with small number of features. Stepwise carries the same idea as subset selection but look at a more restrictive set of models. $2^p >>> p^2, \ where \ k = 2$. Stepwise function of p^2 models generated is less computationally expensive but more importantly hinders the rising concern in over-fitting/ over-training the data.

> Forward & Backward stepwise selection does not look at all possible models "looks only a subset of models" - attractive alternatives to best subset selection.

Forward Stepwise Forward stepwise selection is a feature selection method which has following steps:

[1] M_0 -> null model: Forward stepwise selection begins with a model containing no predictors

[2] Adds predictors to the model one-at-a-time until all of the predictors are in the model.

- At each step the variable that gives the greatest additional improvement to the fit, attributes to lowest RSS = $\sum_{i=1}^{n}(y_i - \hat{y}_i)\}$ is added to the model.
- For $k = 0, ..., p-1$, where the number of observation can be either $n > p \ | n < p$:
  - Consider all $p - k$ models that augment the predictors in M_k with one additional predictor.
  - Choose the best among these $p-k$ models and denote as M_{k+1} which is defined as having lowest RSS or highest R^2.

Backward Stepwise Backward stepwise selection is a feature selection method which has following steps:

[1] M_p -> full model: Forward stepwise selection begins with a model containing all the predictors

[2] For $k = p, p-1, ..., 1$ where requires number of observation $n > p$:

- Consider all k models that contain all but one of the preditors in M_k, for a total of k - 1 predictors.
- Choose the best among these k models, and all it M_{k-1}. Here best is defined as having smallest RSS or highest R^2.
- The exact number of models the backward selection considers is $1 + p(p+1)/2$ models.

Single best subset model selection and Bias on RSS and R^2

The problem:

A graphical RSS or R^2 as a function(p), the curve never gets worse. * As more feature gets added to the linear model + RSS always decreases in monotonic manner + R^2 always increases in monotonic manner

R^2 is a proportions of variance explained by linear regression model and its variance at M_0 {TSS} will never increase by the feature addition. Hence the RSS and R^2 cannot be used to choose among the p-1 models due to their dimensional space difference in size.

The result:

Unlike the subset selection, forward and backward stepwise does not guarantees to find the best possible model out of its all p^2 models (combination of features). Thus there is a likeihood of presence in gap in between these RSS or R^2 curve as a function(p). As it could be that the best model containing k predictors is not a **superset** of the best model ontaining k-1 predictors.

Interestingly, it is this gap and the difference in feature selection by "best subset" and "forward stepwise" can be used to verify the correlation between the features. By means, one can actually get a discrepancy between best subset and forward stepwise procedures during EDA.

Estimating Test Error: Two Approaches

In all of the methods mentioned above its single best model is selected from among $M_0, ..., M_p$ using following metrics.

- Cross-validation prediction error
- C_p/ AIC/ BIC/ Adjusted R^2

The stepwise feature selection has been a popular techniques. However, if this procedure had just been proposed as a statistical methods, it would likely be rejected due to many violation in principle of statistical estimation and hypothesis testing[1]. Here is a summary of the problems with this method.

- Yields R^2 values that are bias high.
- The ordinary F and χ^2 test statistics do not have the claimed distribution
- It is not guranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
- It provides regression coefficients that are biased high in absolute value and need shrinkage.
  - Regression coefficients: will appear larger
  - Confidence intervals: will appear narrower
  - p-values: will appear smaller; also invalid
  - R^2: will appear larger
- Most importantly, this stepwise method is a automatic feature selection procedure which can lead us to not think about the problem caused by collinearity. This is problematic in cases where, for instance, a variable should be definitely included in the model to control for confounding.

  The significance values [a.k.a. p-values] are generally invalid when a stepwise method (stepwise, forward, or bakward) is used. IBM Knowledge Center

Questions

Q:

Can C_p used for only linear model and how in linear model the C_p and AIC are assigned to be the same (proportional relationship)?

- -2logL = RSS/ $sigma^2$
- What exactly is AIC and BIC and how does accounting estimated squared variance of epsilon relates to estimating the test set.
- Are AIC and BIC reliable measure to compare different models? Or would k-fold CV be more acceptable approach to estimate the test error?

Clearer pictures on why stepwise functions neglects the principle in statistical approach.

- Why is the significance values (p-values) are generally invalid when a stepwise function is used?

External Q:

- The i.i.d ideal rule on regression applies to entire family of regression?
- When working with independent variable that is not ~N(mean, sigma) how do we know what distribution (e.g Bernoulli or poisson distribution) it is and how one should treat them?

Reference

[1] Frank E., Harrell. Jr. (2001) *Regression Modeling Strategies With Applications to Linear Models, Logistic Regression, and Survival Analysis* Springer Series in Statistics.