Chapter 6: Linear Model Selection and Regularization

In this chapter, we will consider alternative approaches to the least squares regression method. Alternative fitting procedures can yield better prediction accuracy and model interpretability.

For example, it is often the case that some of the predictors used in a multiple regression model are in fact not associated with the response. Including such irrelevant variables leads to unnecessary complexity and lack of interpretability in the resulting model.

By removing these types of predictors, we can obtain a more accurate predictive model. We will concentrate in the following three important classes of methods:

1. **Subset Selection**: This approach involves identifying the optimal subset of predictors that are related to the response Y. We then fit a model using least squares on the reduced set of variables.
2. **Shrinkage**: This approach involves fitting a model involving all predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimate. This shrinkage (or regularization) reduces the variance term and depending on the type of shrinkage, some regression coefficients will be exactly zero. This implies that regularization methods will also perform a variable selection.
3. **Dimension Reduction**: This method projects the predictors into a smaller number of different linear combinations or projections of these variables. The new subset of predictors is then used to fit a least squares linear regression model.

Note that subset selection, shrinkage and dimension reduction methods can easily be applied to the classification models. In this chapter, we focus only on regression models.

Subset Selection

The following methods are used for selecting subsets of predictors:

- **Best Subset Selection**: We fit a separate least squares regression for each possible combination of the predictors in the data set. The algorithm is described in the steps listed below:

  Step 1: Start with a null model, a model with no predictors. This model simply predicts the sample mean for each observation.

Step 2: Fit all possible models that contain exactly one predictor, all models that contain exactly two predictors, and so forth. Pick the best among all models with one predictor, with two predictors, and so on by looking at the model with the largest R-Squared.

Step 3: Select a single best model from the final candidates using cross-validated prediction error, Mallow's Cp, the Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R-Squared.

Step 2 is performed in the training data and Step 3 is applied in the test data. However, the best selection method requires a lot of computational power, especially for a large number of predictors.

There are attractive alternatives to best subset selection.

- **Forward Stepwise Selection**: The forward stepwise selection begins with a model containing no predictors, and then 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 is added to the model. The basic steps of the forward stepwise selection are listed below:

  Step 1: Start with the null model, which contains no predictors.

  Step 2: Consider all possible models that augment the predictors allowing in one additional predictor at a time. Choose the best among these models with the highest R-Squared.

  Step 3: Select the single best model among all candidates by reviewing the cross-validated prediction error, Cp, AIC, BIC, or adjusted R-Squared on the test data.

- **Backward Stepwise Selection:** Like forward stepwise selection, backward stepwise selection provides an efficient alternative to the best subset selection.

  The backward stepwise selection begins with the full least squares model containing all predictors, and then iteratively removes the least useful predictor, one-at-a-time.

  Cross-validated prediction error, Cp, AIC, BIC or adjusted R-Squared are utilized to choose the best possible model on the test data.

Choosing the Optimal Model

Best subset selection, forward selection, and backward selection result in the creation of a set of models, each of which contains a subset of the p predictors.

If the resulting models are different, we need a decision rule for selecting the best model, or the model with the smallest test error.

There are two approaches to measure the test error. The first indirectly estimates the test error by making an adjustment to the training error to account for the bias due to overfitting. The second approach allows us to directly estimate the test error, using either a validation approach or a cross-validation approach, as described in Chapter 5.

Let's discuss four indirect ways to measure the test error.

Mallow's Cp, given by equation 6.2, gives the estimate of the test MSE for a fitted least squares model. We always choose the model with the smallest Cp.

The AIC criterion is proportional to Cp for least squares models. We would like to choose the model with the smallest AIC coefficient.

Very similarly, we can define the **BIC** criterion in equation 6.3. Just like the case with Cp and AIC, the best model is chosen by the smallest BIC.

The **adjusted R-Squared** approach is another popular approach for selecting the best model among a set of models that contain different number of variables. The adjusted R-squared is calculated in equation 6.4.

A large value of R-Squared indicates a model with a small test error.

In summary, the practitioner needs to carefully examine all these statistics before a final determination is made for the optimal model. Note that the formulas for Cp, AIC and BIC are presented in the case of a linear model fit using least squares. These quantities, however, can also be defined for more general types of models.

Shrinkage Methods

We can fit a model containing all p predictors using a technique that constrains or regularizes the coefficient estimates, by shrinking the regression coefficients toward zero.

There are two very popular techniques to achieve shrinkage. The first technique is ridge regression and the second is lasso.

Ridge Regression

Ridge regression is very similar to the least squares regression, except that the coefficients are estimated by minimizing a slightly different quantity given by equation 6.5.

This equation includes a tuning parameter, lambda, which is to be determined separately. There is also the second term of the equation, the shrinkage penalty and it has the effect of shrinking the regression coefficients toward zero.

When the tuning parameter is zero, the penalty term has no effect and ridge regression will produce the least square estimates.

However, as the tuning parameter increases and the impact of the shrinkage penalty grows, and the ridge regression coefficients will approach zero.

Selecting a good value for lambda is critical and it involves an iterative process. Different values of lambda will a generate different set of ridge regression coefficients.

The main advantage of ridge regression over least squares is rooted in the bias-variance trade-off. As the tuning parameter increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but possible increase in bias.

Ridge regressions work the best when the least squares regression coefficients have high variance. Ridge regressions also have a computational advantage over the best subset selection.

The Lasso

Ridge regression has one major disadvantage: the inclusion of all predictors in the final model. The penalty term will shrink all of the coefficients toward zero, but it will not set any of them exactly to zero.

The lasso method is an alternative to ridge regression that overcomes this advantage. By examining the lasso function 6.7, we see that both methods have similar formulations.

The only difference is that the ridge regression penalty has been replaced by the lasso penalty.

The lasso penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter is sufficiently large.

The lasso performs a variable selection. Selecting a good value for the tuning parameter is critical!

Neither the lasso nor the ridge regression will universally dominate the other. Cross-validation can be used to determine which approach is better on a particular data set.

Implementing ridge regression and lasso requires the following method of choosing a value for the tuning parameter:

- Choose a grid of lambda values and compute the cross-validation error for each value of lambda.
- We then choose the tuning parameter value with the smallest cross-validation error.
- Finally, the model is re-fit using all of the available observations and the selected value of the tuning parameter.

Dimension Reduction Methods

Dimension reduction methods transform the predictors and then fit a least squares model using the transformed variables.

Linear combinations of the original predictors are considered as shown in equation 6.16 and we can then apply the least squares fit.

All dimension reduction methods work in two steps. In the first step, transformed predictors are obtained. In the second step, the model is fit using the new set of predictors.

We can select the new set of predictors by considering two approaches: principal components and partial least squares.

Principal Components Regression

Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables.

PCA is also used as a tool for unsupervised learning. The first principal component direction of the data is that along which the observations vary the most.

The principal components regression (PCR) approach involves constructing the first M principal components and then using these components as predictors in a linear regression model that is fit using least squares.

The key idea is that often a small number of principal components is enough to explain most of the data variability. That's where PCR performs extremely well!

We choose the number of principal components by cross-validation.

Partial Least Squares

Note that the PCR approach involves identifying linear combinations or directions that best represent all predictors. These directions are derived in an unsupervised way, since the response Y is not used at all in determining the principal components.

This is a clear disadvantage for PCR: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

A supervised alternative to PCR is the partial least squares method (PLS).

PLS makes use of the response Y in order to identify new predictors, which are not only approximating the old ones but also are related to the response. The PLS approach attempts to find directions that help explain both the response and the predictors.

PLS computes the first data direction by setting each coefficient equal to the simple linear regression of Y onto a given predictor. In other words, PLS places the highest weight on the variables that are most strongly related to the response.

Assignments for class discussion

Lab 1: Subset Selection Methods

Lab 2: Ridge Regression and Lasso

Lab 3: PCR and PLS Regression