Lecture 101

Date: 27 March 2020
Scribes: Akash Gupta
Topics: Model selection and Regularization

0.1 Shrinkage

Shrinkage is a process by which we fit all the p variables of the model however the coefficient estimates are effectively shrunken towards 0 when compared to the least squares coefficient values. This shrinkage process is called **regularization** and it helps in reducing variance of the estimates. Note that some coeffients might also be shrunk to 0 and hence effectively performing variable selection on the model. Remember that variable selection is a way by which the predictor variables that have little or no impact on the response are removed from the model.

0.1.1 Ridge regression

Remember that in least squares regression, we essentially select coefficients such that our RSS is minimized :-

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) \tag{1}$$

In ridge regression however we minimize a slightly different quantity, which is given by $\hat{\beta}^R$ as we shall see later. So the quantity we minimize is :-

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

The term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called the shrinking penalty and it this term that actually reduces the coefficient estimates towards 0. Remember that if $\lambda = 0$ then there is no shrinking taking place and we original least squares estimates are obtained. However, if $\lambda \to \infty$ then the shrinking penatly term becomes more impactful and hence the coefficients tend towards 0 in this situation. Depending on the critical value of λ that we select we will obtain rather different coefficient estimates of regression in the form of $\hat{\beta}_\lambda^R$. Note that we do not apply shrinkage to the intercept and remember that **we want to shrink the estimated association of the predictors with the response**. If the columns of data matrix $\boldsymbol{X}$ have been centered with mean 0 then we will get intercept as :-

$$\hat{\beta}_0 = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{3}$$

Note that centering means transforming variables to standard normal form. Remember the main point that **as λ increases the ridge coefficient estimate tends**

to 0.

To get an idea of the relative amount that the coefficient estimates have been shrunken to 0 we compute ratio of the norm of the ridge coefficient vector and the lease squares coefficient vector. The formula for computing norm and the conditions are given by :-

$$||\hat{\beta}|| = \sqrt{\sum_{j=1}^{p} \beta_j^2} \tag{4}$$

$$\frac{||\hat{\beta}_\lambda^R||}{||\hat{\beta}||} = 0, \lambda \to \infty \tag{5}$$

$$\frac{||\hat{\beta}_\lambda^R||}{||\hat{\beta}||} = 1, \lambda \to 0 \tag{6}$$

An important property is that in least squares case, the coefficient values do not depend on scaling of the predictor variables. However, in case of ridge coefficients, a different scaling pattern among different predictors is likely to have an impact on the values of the ridge coefficients and that is precisely why we standardize the predictor variables to standard normal values with mean 0 and variance 1 as follows :-

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{ij})^2}} \tag{7}$$

The main point that the ridge regression achieves is that it changes the bias-variance trade off of our model. As we increase λ the flexibility of the model decreases and as a result, variance decreases and bias increases. So we essentially trading off for lesser variance at a cost of higher bias. As we increase λ initially till about a value of 30 the variance decreases quite a lot while the corresponding increase in bias is not so much, and hence it as at this point where our MSE is minimum. Beyong this value of λ, for every unit decrease in variance, the bias increase is quite a lot, making the MSE also higher.

0.1.2 Lasso

One problem of the ridge regression is that while it does improve our model and is better than subset selection techniques, it still fits all the variables in the model, albeit shrinking them towards 0. The lasso is better in the sense that it can additionally even remove the seemingly irrelevant variables from the model for better interpretation. The lasso coefficients given by $\hat{\beta}_\lambda^L$ and they are selected so as to minimize the expression :-

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{8}$$

$$||\hat{\beta}|| = \sum_{j=1}^{p}|\beta_j| \tag{9}$$

Previoulsy we used the l_2 norm and now we use the l_1 norm given in equation 9. In equation 8 the term to the right of RSS is the 'Lasso penalty' term. The primary difference between ridge and lasso is that the lasso penalty forcees certain coefficients to exactly 0 while performing shrinkage operation as $\lambda \to \infty$. Hence in effect it is also performing variable selection by removing variables.

0.1.3 Logic behind the lambda terms

The equations of minimizing ridge and lasso regressions can be thought of as derived from lagrangian methods where we have an objective function and constraints. For Ridge and Lasso respectively it is given by :-

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right), subject\ to\ \sum_{j=1}^{p} \beta_j^2 \leq s \tag{10}$$

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right), subject\ to\ \sum_{j=1}^{p} |\beta_j| \leq s \tag{11}$$

Therefore, for every value of λ in the regular equations, there is a value of s such that as long as the constraint is satisfied, these new equations also give the same coefficient values. In a way, when we perform Lasso, then we are **trying to find coefficient estimates that minimize RSS subject to the constraint that there is a budget s that tells us how large $\sum_{j=1}^{p} |\beta_j|$ can be** . Note that if the budget is large enough, then even the coefficients will be large and we will just obtain the least squares estimates in that case.

We should ideally perform ridge and lasso regressions on models to improve their fit and make their variability less. Consequently the lasso or ridge version can be chosen as the final model only by performing cross validation techniques on it - that is basically computing the test error rates.

0.1.4 Selecting tuning parameter

While selecting the optimal value of the tuning parametere, we take a grid of λ values and compute the cross validation test error rate for each λ. Then we choose the model and tuning parameter value that corresponds to the lowest CV error rate.