STAT 383C: Statistical Modeling I **Fall 2016**

Lecture 10 — September 27

Lecturer: Purnamrita Sarkar *Scribe: Amin Anvari*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

10.1 Regularization

Regularization methods such as Ridge Regression and LASSO introduce a penalty term on model complexity to prevent overfitting.

10.1.1 Ridge Regression

For least squares regression, the model coefficients are selected by

$$\hat{\boldsymbol{\beta}}_{LS} = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}_0 - x_i^T \boldsymbol{\beta})^2 \tag{10.1}$$

For ridge regression, an additional term is added which penalizes all $\boldsymbol{\beta}_j$ for $j > 0$

$$\hat{\boldsymbol{\beta}}_{ridge} = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} ((y_i - \boldsymbol{\beta}_0 - x_i^T \boldsymbol{\beta})^2 + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \tag{10.2}$$

It is clear from this equation that if the variables are on different scales the Ridge Regression model will penalize them differently. Thus, centering must be performed to remove the means from all parameters. Additionally, the standard errors must be normalized to 1.

First lets forget about the normalization, and just consider the original problem. We will see that centering $\boldsymbol{X}$ and $\boldsymbol{y}$ ($X_c(i,j) = X_{ij} - \bar{x}_j$ and $y_c(i) = y_i - \bar{y}$ will not change the β_i's other than the intercept term.

$$\begin{aligned} &\sum_i (y_i - \beta_0 - \sum_j \beta_j x_{ij})^2 + \lambda \beta_j^2 \\ &= \sum_i (y_i - (\beta_0 + \sum_j \beta_j \bar{x}_j) + \sum_j \beta_j (x_{ij} - \bar{x}_j))^2 + \lambda \beta_j^2 \end{aligned}$$

$$\boldsymbol{\beta}'_0 = \boldsymbol{\beta}_0 + \sum_{j=1}^{k} x_{ij} \boldsymbol{\beta}_j \tag{10.3}$$

$$\boldsymbol{\beta}'_j = \boldsymbol{\beta}_j, \forall j > 0 \tag{10.4}$$

As it turns out, deriving the above expression w.r.t $\boldsymbol{\beta}'_0$ and setting to zero gives:

$$\hat{\boldsymbol{\beta}'}_0 = \bar{y}$$

Thus the reparameterized equation for $\hat{\boldsymbol{\beta}}_{ridge}$ is

$$OBJ_{ridge} = \min_{\boldsymbol{\beta}} (\boldsymbol{y}_c - \boldsymbol{X}_c\boldsymbol{\beta})^\top (\boldsymbol{y}_c - \boldsymbol{X}_c\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}, \tag{10.5}$$

where $\boldsymbol{y}_c$ and $\boldsymbol{X}_c$ are centered versions of $\boldsymbol{y}$ and $\boldsymbol{X}$. Setting the derivative of this expression to zero gives the solution

$$\hat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^\top y \tag{10.6}$$

Note that this ridge regularization is the same as adding λ to all eigenvalues of $\boldsymbol{X}^T X$. Furthermore, exactly corresponding to the hat matrix, in this case if we write $\hat{y} = Sy$ we get:

$$S = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I)^{-1} \boldsymbol{X}^\top \tag{10.7}$$

and so the effective degrees of freedom is:

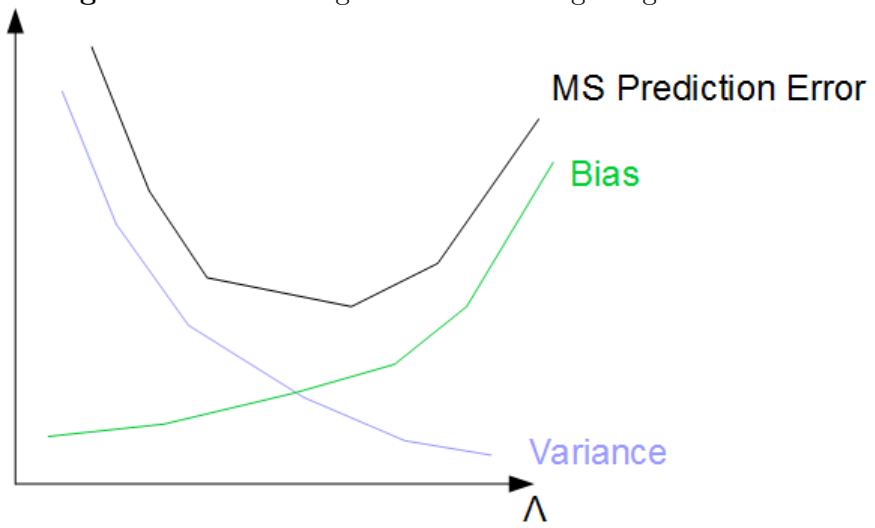
$$\text{trace}(S) = \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \tag{10.8}$$

In which σ_i's are the singular values of matrix $\boldsymbol{X}$.

How to Choose λ: Selection of λ is a tradeoff between bias, variance and mean square error (Fig. 9.3).

1. Perform a grid search over λ or $log(\lambda)$.
2. For each λ perform a cross-validation and calculate the mean and standard error on the estimated model prediction error
3. Identify the λ with the lowest mean predictive error
4. Apply One Standard Error Rule to select most parsimonious model whose mean lies within one standard error (in this case more parsimonious means smaller effective degrees of freedom, so a larger λ)

Figure 10.1. Selecting Lambda For Ridge Regression Model



Equivalent Bayesian Interpretation Assume that y is normally distributed, and apply a Gaussian prior to $\boldsymbol{\beta}$

$$y \sim N(\boldsymbol{X\beta}, \sigma^2 I) \tag{10.9}$$
$$\boldsymbol{\beta} \sim N(0, \tau^2 I) \tag{10.10}$$

Then the posterior distribution of $\boldsymbol{\beta}$ given y can be calculated as

$$f(\boldsymbol{\beta}|y) = \exp(\frac{-(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta})}{2\sigma^2})\exp(\frac{-\boldsymbol{\beta}^\top\boldsymbol{\beta}}{\tau^2}) \tag{10.11}$$

From this form it can be observed that the relationship between σ and τ defines λ. Ridge regression generally has smaller LSE than linear regression, and Ridge Regression will make coefficients small for variables which are not highly correlated to the output. However, Ridge Regression will not drive coefficients to zero (unless $\lambda = \infty$, in which case it drives all coefficients to zero), so it cannot be used for variable selection.

10.1.2 LASSO

$$OBJ_{LASSO} = \min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda||\boldsymbol{\beta}||_1 \tag{10.12}$$

The LSE of LASSO is comparable to Ridge Regression, and LASSO will drive some coefficients to zero with a large λ. However, there is no closed form solution. If the problem is reformulated as

$$\hat{\boldsymbol{\beta}}_{LASSO} = \min_{\boldsymbol{\beta}}(\boldsymbol{y}-\boldsymbol{X\beta})^\top(\boldsymbol{y}-\boldsymbol{X\beta}) \tag{10.13}$$

$$\text{s.t.} \sum_{j=1}^{k}|\boldsymbol{\beta}_j| \leq t \tag{10.14}$$

This is a convex optimization problem, which will have a solution even without a closed-form.

Note that we can also devise an objective function using the zero-norm instead of one-norm (All Subsets objective function):

$$OBJ_{\text{All subsets}} = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) s.t. \|\boldsymbol{\beta}\|_0 \leq k \tag{10.15}$$

The problem with this objective function is that it does not yield a convex problem anymore. In fact, one would need to search through all possible subsets of size smaller than $k+1$ to evaluate the best possible subset.

Consider the easier problem where I want the best subset of size k. We will denote this by OBJ_{BS} from now on. Typically this is a NP complete problem, which is computationally intractable. However, as it turns out in the case where the design matrix has orthogonal columns, i.e. the features are orthonormal to each other, i.e. $\boldsymbol{X}^T\boldsymbol{X} = I$ one can actually write down effect of best subset of size k, L_1 and L_2 norm penalizations easily.

10.2 Orthogonal design matrix

Consider an orthogonal design matrix such that $\boldsymbol{X}^T\boldsymbol{X} = I$. Now we will look at Best subset of size k, Lasso and Ridge to get an intuition about how the different regularizations work. The main trick is to see that in this case the loss function becomes separable w.r.t the coefficients of $\boldsymbol{\beta}$. Also note that in this case $\boldsymbol{\beta}^{ols} = \boldsymbol{X}^T\boldsymbol{y}$. So

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{\beta} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{\beta}^{ols} - \boldsymbol{\beta})^T(\boldsymbol{\beta}^{ols} - \boldsymbol{\beta}) + \boldsymbol{y}^T(I - \boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{y}$$

Since the last term is independent of $\boldsymbol{\beta}$, we only need to consider the first term.

10.2.1 Best subset of size k

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{10.16}$$

$$\text{s.t. } \|\boldsymbol{\beta}\|_0 = k \tag{10.17}$$

which is equivalent to:

$$\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \boldsymbol{\beta}^{ols})^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{ols}) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{p} (\beta_i - \beta_i^{ols})^2 \tag{10.18}$$

$$\text{s.t. } ||\boldsymbol{\beta}||_0 = k \tag{10.19}$$

Now think for a second. I can only put nonzero values in k positions, and at the same time minimize the above sum of squares. Can I put them on the smallest k β_i values? That

will incur a lot of error from the large coefficients. So, we should use:

$$\hat{\beta_i^{BS}} = \begin{cases} \hat{\beta}_i^{OLS} & \text{If } |\beta_i^{ols}| \geq |\boldsymbol{\beta}_{(k)}^{ols}| \\ 0 & \text{Otherwise} \end{cases}$$

Here $x^{(k)}$ means the k^{th} largest element in absolute value.

10.2.2 Lasso

Now lets think about Lasso. Now the objective function is

$$\min_{\boldsymbol{\beta}} \sum_i (\beta_i - \beta_i^{ols})^2 + \lambda \sum_i |\beta_i| \tag{10.20}$$

$$\tag{10.21}$$

Since the function is now separable w.r.t the elements, we can reason about each coordinate separately. Lets take coordinate i. If $\beta_i^{ols} > 0$, it makes sense to set $\beta_i > 0$ in order to minimize the above function. So that gives us $\min_{\beta_i}(\beta_i^{ols} - \beta_i)^2 + \lambda\beta_i$. Derive w.r.t β_i and set to zero to solve. $\beta_i^{LASSO} = \beta_i^{ols} - \lambda/2$, but in this case $\beta_i^{LASSO} > 0$ as well. So, use: $\beta_i^{LASSO} = \max(\beta_i^{ols} - \lambda/2, 0)$. A similar argument for the negative OLS coefficients give the following:

$$for \qquad \beta_i^{ols} > 0: \qquad \hat{\beta}_i = \max\{\beta_i^{ols} - \frac{\lambda}{2}, 0\} \tag{10.22}$$

$$for \qquad \beta_i^{ols} \leq 0: \qquad \hat{\beta}_i = \min\{\beta_i^{ols} + \frac{\lambda}{2}, 0\} \tag{10.23}$$

10.2.3 Ridge

Finally, lets figure out ridge regression. This is actually quite straight forward. We want:

$$\min_{\boldsymbol{\beta}} \sum_i (\beta_i - \beta_i^{ols})^2 + \lambda \sum_i \beta_i^2 \tag{10.24}$$

Derive both sides w.r.t β_i and set to zero. This gives:

$$\hat{\beta}_i = \frac{\beta_i^{ols}}{1 + \lambda}$$