Lab Solutions 07

1. This was a question from last lab, but in case you did not get there, make sure that you can derive a formula for ridge regression under the uncorrelated assumption.

In the previous lab, I assumed that X^tX was equal to n times the identity. Removing the n term gives:

$$\beta^{RIDGE} = \frac{1}{1+\lambda} \cdot X^t y = \frac{1}{1+\lambda} \cdot \beta^{OLS}.$$

See the previous lab solutions for the details.

2. Write the best subset selection loss function as a sum over j from $j = 1$ to $j = p$. That is, you can write the loss function as a sum of independent terms, where each elements depends only on the j'th column of X and the j'th component of β.

Writing $\chi\{b_j \neq 0\}$ as the indicator function that is one when the statement is true and zero otherwise, we have:

$$\begin{aligned}
||y - Xb||_2 + \lambda ||b||_0 &= y^t y + b^t X^t X b - 2y^t X b + \lambda \cdot \sum_j \chi\{b_j \neq 0\} \\
&= y^t y + b^t b - 2y^t X b + \lambda \cdot \sum_j \chi\{b_j \neq 0\} \\
&= y^t y + \sum_j b_j^2 - 2 \sum_j y^t X_j b_j + \lambda \cdot \sum_j \chi\{b_j \neq 0\} \\
&= y^t y + \sum_j \left[b_j^2 - 2y^t X_j b_j + \lambda \cdot \chi\{b_j \neq 0\} \right]
\end{aligned}$$

Which is now a decoupled sum over the components of β_j, plus a leading constant term that doesn't effect the loss.

3. Repeat the previous question for lasso regression, showing that it also decouples over the individual variables.

The lasso regression works similarly:

$$
\begin{aligned}
||y - Xb||_2 + 2\lambda ||b||_1 &= y^t y + b^t X^t X b - 2y^t X b + 2\lambda \cdot \sum_j |b_j| \\
&= y^t y + b^t b - 2y^t X b + 2\lambda \cdot \sum_j |b_j| \\
&= y^t y + \sum_j b_j^2 - 2\sum_j y^t X_j b_j + 2\lambda \cdot \sum_j |b_j| \\
&= y^t y + \sum_j \left[b_j^2 - 2y^t X_j b_j + 2\lambda \cdot |b_j| \right]
\end{aligned}
$$

The only difference being the last term in the equation.

4. Understand why your answers to the last two questions allow us to minimize the loss function individually for each component β_j.

In general if we want to minimize a function $f(x, y) = f_1(x) + f_2(y)$ over x and y, it should be clear that all we need to do is minimize f_1 over x and f_2 over y. The same logic applies to any number of variables.

5. Assume that $\beta_j \neq 0$. What would be its optimal value under best subset regression? When is the loss at this point better than setting $\beta_j = 0$? Put these conditions together to get a general formula for β_j under best subset selection.

We want to minimize the quantity:

$$
f_j(b_j) = b_j^2 - 2y^t X_j b_j + \lambda \cdot \chi\{b_j \neq 0\}
$$

If $\beta_j \neq 0$ then, this is equal to a simple quadratic function in b_j:

$$
b_j^2 - 2y^t X_j b_j + \lambda
$$

We minimize this just like any other quadratic, by taking the derivative with respect to b_j and setting it equal to zero:

$$
\begin{aligned}
2b_j - 2y^t X_j &= 0 \\
b_j &= y^t X_j
\end{aligned}
$$

The value of f_j at this point is:

$$
\begin{aligned}
f_j(y^t X_j) &= (y^t X_j)^2 - 2(y^t X_j) \cdot (y^t X_j) + \lambda \\
&= \lambda - (y^t X_j)^2
\end{aligned}
$$

The question is, when will this be better than the alternative of setting $b_j = 0$? There we have $f_j(0) = 0$, which will be better – more negative – whenever $\lambda \geq (y^t X_j)^2$. This should seem reasonably because we are more likely to set a coefficient equal to zero if (i) λ is large, or (ii) the correlation $y^t X_j$ is small.

Putting this together, we can compactly write:

$$
\beta^{BSR} = \begin{cases} 0 & \lambda > (y^t X_j)^2 \\ y^t X_j & \text{else} \end{cases}
$$

Or, even better, as:

$$
\beta^{BSR} = \begin{cases} 0 & \lambda > (\beta_j^{OLS})^2 \\ \beta_j^{OLS} & \text{else} \end{cases}
$$

Which shows the direct link with the unpenalized solution.

6. Use a similar approach to find a solution for β_j for lasso regression. This requires a little bit more work but is very doable with simple one-dimensional calculus!

We want to minimize the quantity:

$$
f_j(b_j) = b_j^2 - 2y^t X_j b_j + 2\lambda \cdot |b_j|
$$

Assume to start that the optimal value of b_j is greater than zero. Then, $|b_j| = b_j$ and we only have to consider points of f_j that can be written as a simply quadratic function:

$$f_j(b_j) = b_j^2 - 2y^t X_j b_j + 2\lambda b_j$$

The optimal value occurs when the derivative is equal to zero, which happens at:

$$\begin{aligned} f_j'(b_j) &= 2b_j - 2y^t X_j + 2\lambda \\ 0 &= b_j - y^t X_j + \lambda \\ b_j &= y^t X_j - \lambda. \end{aligned}$$

Great! Except, this is only valid when $b_j > 0$. Therefore, this solution only works when $y^t X_j > \lambda$. Now, assume instead that f_j is optimized when $b_j < 0$. Now it reduces to:

$$f_j(b_j) = b_j^2 - 2y^t X_j b_j - 2\lambda b_j$$

And setting the derivative equal to zero yields:

$$\begin{aligned} f_j'(b_j) &= 2b_j - 2y^t X_j - 2\lambda \\ 0 &= b_j - y^t X_j - \lambda \\ b_j &= y^t X_j + \lambda. \end{aligned}$$

This, in turn, only holds if $y^t X_j \leq -\lambda$. So, what happens if $|y^t X_j| \leq \lambda$? Neither of these conditions hold, and f_j is optimized when $b_j = 0$. Putting this all together yields:

$$\beta^{LASSO} = \begin{cases} y^t X_j + \lambda & y^t X_j \leq -\lambda \\ 0 & |y^t X_j| \leq \lambda \\ y^t X_j - \lambda & y^t X_j \geq \lambda \end{cases}$$

Or, very compactly, as:

$$\beta^{LASSO} = \begin{cases} \beta_j^{OLS} - \lambda \cdot sign(\beta_j^{OLS}) & |y^t X_j| \geq \lambda \\ 0 & \text{else} \end{cases}$$

In words, the ordinary least squares coefficients are shrunk towards zero by a linear factor of λ. Coefficients with absolute size less than λ are set to zero.

7-9. Draw a sketches for the three estimators

Look at the R output for an example of the ridge and lasso estimators. The plot of the best-subset regression is straightforward if you have the correct equation; it is a piecewise constant function.