Handout 06: Ridge Regression

On the last handout I asked to consider a regression problem with an estimated value for β and a data matrix X factorized using the SVD as UDV^t. Then, we considered the predictions from a new $\tilde{\beta}$ equal to β plus a multiple of the smallest right singular vector (V_p). This is given by:

$$\begin{aligned} X(\tilde{\beta}) &= X(\beta + aV_p) & (6.1)\\ &= X\beta + aXV_p & (6.2)\\ &= X\beta + a\sigma_p. & (6.3) \end{aligned}$$

In the lab questions, you assumed $\sigma_p = 0$ and this shows that the predictions $\widehat{y}$ for β are exactly equivalent to the predictions for $\tilde{\beta}$. What if σ_p is positive but small? In this case the predictions are not exactly the same but they are still very difficult to distinguish. Under sufficient noise it is still nearly impossible to distinguish between these two solutions when σ_p is small. This can make regression very difficult to perform because large datasets often have a smallest singular value that is quite small (more on this later).

The fundamental problem here is that we are only the mean squared error as our loss function. Therefore, there is no easy way to distinguish between using β and $\tilde{\beta}$. One solution is to modify the loss function to make it easier to distinguish between these two solutions. For example, here is the equation for ridge regression:

$$\beta_\lambda = \arg\min_b \left\{ ||y - Xb||_2^2 + \lambda ||b||_2^2 \right\} \tag{6.4}$$

For some constant $\lambda > 0$. It says that you want to minimize the errors in prediction, but with an additional cost associated with large values of β. This helps to distinguish between the many possible models and often does a much better job than the ordinary least squares estimator at predicting future values. You can derive a very elegant solution to the ridge regression, particular when you incorporate the SVD. I will have you derive this in the following lab questions.

LAB QUESTIONS

1. Take the gradient of the ridge regression loss function and set it equal to zero. Get an equation for β_λ in terms of X and y. You may need to use the fact that $b^t b$ is equal to $b^t I_p b$ in this derivation. (Note: Do **not** yet use the SVD of X here).

2. The eigenvalue decomposition of a matrix writes a matrix as $Q^t \Lambda Q$ for a diagonal matrix Λ (the entries are called the matrix eigenvalues) and an orthogonal matrix Q. Unlike the SVD, the eigenvalue decomposition only applies to square matrices and even then does not always exist. Show that the eigenvalues of $X^t X$ are equal to the squared singular values of X.

3. I want to derive a formula for β_λ in terms of the SVD of X; this is not a long derivation but does require a trick. Take the equation that you have already derived for β_λ; there should be an identity matrix in the formula. consider the SVD of X as UDV^t. Write the identify matrix in the equation as $V^t V$, plug in the SVD for X, and simplify. You should be able to factor out some of the terms and are left with something very similar to equation we had for the ordinary leasts squares estimator.

4. Understand that the ridge regression is equivalent to fitting ordinary least squares on a new matrix $\bar{X}$ where the singular values have been increased by a factor of λ. Given our argument about the problems with the smallest singular value, does it make sense that this change alleviates the problem of identifiability?

5. Let's somewhat switch gears here and consider a specific example problem. Let $p = 2$ and assume that the first column of X (X_1) can be written as:

$$X_1 = \alpha + X_2, \quad \alpha \in \mathbb{R}^n \tag{6.5}$$

Where α is a small noise vector. So, X_1 and X_2 are very similar to one another. Write an equation for the value Xb, factoring in terms of α and X_2 (there should not be any X_1 left in the equation). Then, assume that we have data generated by:

$$y = X_2 + \text{noise} \tag{6.6}$$

Where the noise is not too large. Convince yourself that all of the following values of b produce a reasonable estimate for $\widehat{y} = Xb$:

$$b = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{6.7}$$

$$b = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{6.8}$$

$$b = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \tag{6.9}$$

$$b = \begin{bmatrix} -100 \\ 101 \end{bmatrix} \tag{6.10}$$

What do you think is the approximate value of β_λ for ridge regression for a small value of λ assuming the noise vector and α are also both small?

6. For the previous question, can you guess a plausible value for V_p (the last left singular) of the matrix X? Look back at handout 5, question 3, for a hint.

7. In the following final questions, we are going to consider a dataset where:

$$X^t X = 1_p \cdot n. \tag{6.11}$$

   There is nothing you need to compute here, but make sure that you understand: (a) why it makes sense to include the n on the right-hand side, and (b) why it makes sense that we say in this case that the columns of X are uncorrelated.

8. Taking $X^t X = 1_p \cdot n$, what is the value of the ordinary least squares estimator? Can you explain exactly what a particular component β_k is in terms of an inner product?

9. Again taking $X^t X = 1_p \cdot n$, write the ridge regression vector β_λ as a function of just n, λ, and the OLS solution β.