Regularization

compiled by Alvin Wan from Professor Benjamin Recht's lecture

Recall that for any problem with prediction error, we have the following trade-off:

- With finite data, we have more variance.
- With a simpler model, we have more bias.

How can we interpolate the two, so that we minimize the costs of both? We can start off with a simple example, using least squares.

1 Least Squares

Consider the following model for our data, where X is $n \times d$, $\vec{y}$ is $n \times 1$, and β is $d \times 1$.

$$\begin{aligned} X &\sim \mathcal{N}(0, \beta^2 I) \\ \vec{y} &= X\beta + e \end{aligned}$$

We have that our noise $e = [\epsilon_1 \cdots \epsilon_N]^T$ is a vector of normally distributed $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let us start by restating the least squares objective function. Note that x_i is the ith sample and ith row from X. *However, per convention*, we will consider each x_i to be a column vector.

$$\text{minimize} \sum_{i=1}^{n} (x_i^T \beta - y_i)^2$$

Now, we consider its optimal solution $\hat{\beta}$.

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} (X^T (Xv + e)) \\ &= (X^T X)^{-1} (X^T X v + X^T e) \\ &= \beta + (X^T X)^{-1} X^T e \end{aligned}$$

1.1 Bias of Least Squares

We will now compute its bias, which we find to be 0. Now, consider some sample x (a $d \times 1$ column vector) and corresponding label $y = x^T\beta + \epsilon$, a sample and label from our validation set.

$$\mathbb{E}[\hat{y} - y] = \mathbb{E}[x_i^T\hat{\beta} - y]$$

First, plug in $\hat{\beta} = \beta + (X^TX)^{-1}X^Te$.

$$\begin{aligned}\mathbb{E}[x^T\hat{\beta} - y] &= \mathbb{E}[x^T(\beta + (X^TX)^{-1}X^Te) - y] \\ &= \mathbb{E}[x^T\beta + x^T(X^TX)^{-1}X^Te - y]\end{aligned}$$

Now, plug in what we have for $y = X\beta + e$. Below, we use the fact that $(X^TX)^{-1} = X^{-1}X^{-T}$. In other words, take the inverse of each matrix and swap the order.

$$\begin{aligned}\mathbb{E}[x^T\beta + x^T(X^TX)^{-1}X^Te - y] &= \mathbb{E}[x^T\beta + x^T(X^TX)^{-1}X^Te - (x\beta + \epsilon)] \\ &= \mathbb{E}[x^T\beta + x^T(X^TX)^{-1}X^Te - x\beta - \epsilon] \\ &= \mathbb{E}[x^T(X^TX)^{-1}X^Te - \epsilon]\end{aligned}$$

First, by linearity of expectation and the fact that ϵ is normally distributed around 0, we can apply the following.

$$\begin{aligned}\mathbb{E}[x^T(X^TX)^{-1}X^Te - \epsilon] &= \mathbb{E}[x^T(X^TX)^{-1}X^Te] - \mathbb{E}[\epsilon] \\ &= \mathbb{E}[x^T(X^TX)^{-1}X^Te]\end{aligned}$$

We can consider $M = (X^TX)^{-1}X^Te$ to be some $d \times 1$ matrix where each entry is a product of i.i.d. normally-distributed random variables around 0.

$$\mathbb{E}[x^T(X^TX)^{-1}X^Te] = 0$$

1.2 Variance of Least Squares

We will now compute the variance of $\text{var}(\hat{y} - y)$. Using the previous part, we know that $\hat{y} - y$ reduces to $x^T(X^TX)^{-1}X^Te$.

$$\text{var}(\hat{y} - y) = \text{var}(x^T(X^TX)^{-1}X^Te)$$

Remember that the definition of variance states $\text{var}(\hat{y} - y) = \mathbb{E}[(\hat{y} - y)^2] - \mathbb{E}[\hat{y} - y]^2$. Earlier, we showed that $\mathbb{E}[\hat{y} - y] = 0$, so $\text{var}(\hat{y} - y) = \mathbb{E}[(\hat{y} - y)^2]$.

$$\begin{aligned}
\text{var}(\hat{y} - y) &= \mathbb{E}[(\hat{y} - y)^2] \\
&= \mathbb{E}[(x^T(X^TX)^{-1}X^Te)^2]
\end{aligned}$$

Since $\hat{y} - y$ is a scalar, we can represent $(\hat{y} - y)^2$ as $(\hat{y} - y)(\hat{y} - y)^T$. (Remember that the typical translation of the power of two is $(\hat{y} - y)^T(\hat{y} - y)$ so that the dot product yields a scalar.) We also use the fact that $(X^TX)^T = X^TX$.

$$\begin{aligned}
\mathbb{E}[(x^T(X^TX)^{-1}X^Te)^2] &= \mathbb{E}[(x^T(X^TX)^{-1}X^Te)(x^T(X^TX)^{-1}X^Te)^T] \\
&= \mathbb{E}[x^T(X^TX)^{-1}X^Tee^TX(X^TX)^{-T}x] \\
&= \mathbb{E}[x^T(X^TX)^{-1}X^Tee^TX(X^TX)^{-1}x]
\end{aligned}$$

Given $\text{var}(e) = \mathbb{E}[e^2] - \mathbb{E}[e]^2$. Since $e \sim \mathcal{N}(0, \sigma^2)$, $\mathbb{E}[e]^2 = 0$ making $\text{var}(e) = \mathbb{E}[e^2] = \sigma^2$. We additionally know that e and ee^T are independent of all other X and x. By independence, we can then rewrite:

$$\begin{aligned}
\mathbb{E}[x^T(X^TX)^{-1}X^Tee^TX(X^TX)^{-1}x] &= \mathbb{E}[ee^T]\mathbb{E}[x(X^TX)^{-1}X^TX(X^TX)^{-1}x] \\
&= \sigma^2\mathbb{E}[x^T(X^TX)^{-1}X^TX(X^TX)^{-1}x] \\
&= \sigma^2\mathbb{E}[x^T(X^TX)^{-1}x]
\end{aligned}$$

We take $X^T X \approx n\beta^2 I$. Note that since $\mathbb{E}[x_i] = 0$, $\text{var}(x_i) = x_i^2$. This implies that $\mathbb{E}[x^T x] = \mathbb{E}[\|x\|^2] = d\sigma^2$. Since $X \sim \mathcal{N}(0, \beta^2 I)$, $\mathbb{E}[x^T x] = d\beta^2$.

$$\begin{aligned}
\sigma^2 \mathbb{E}[x^T (X^T X)^{-1} x] &\approx \sigma^2 \mathbb{E}[x^T (n\beta^2 I)^{-1} x] \\
&= \frac{\sigma^2}{n\beta^2} \mathbb{E}[x^T x] \\
&= \frac{\sigma^2}{n\beta^2} (d\beta^2) \\
&= \frac{\sigma^2 d}{n}
\end{aligned}$$

Our variance is thus approximately $\frac{\sigma^2 d}{n}$. We see that variance thus increases with the number of features d but decreases with the number of sample points n. Thus, variance is proportional to the complexity of our model and is inversely related to the amount of data.

2 Ridge Regression

We now consider another minimization problem, effectively least-squares but with a l2-norm penalty. We begin by restating the objective function, first in terms of individual samples and then in matrix form.

$$\text{minimize}_w \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2$$

$$\text{minimize}_w \frac{1}{n} \|Xw - y\|^2 + \lambda \|w\|^2$$

We know that the solution is the following.

$$\hat{w} = (X^T X + n\lambda I)^{-1} X^T y$$

In the above, since $X^T X$ is always positive semidefinite, $X^T X + n\lambda I$, where $n > 0$ ensures that $X^T X + n\lambda I$ is positive definite and thus invertible.

We now make the following claim: As $\lambda \to \infty$, variance goes to 0 and bias increases.

2.1 Bias of Ridge Regression

Consider the following example, where as above, we assume $\frac{1}{n}X^TX \approx \beta^2 I$. We can then pre-multiply the objective function by X^T. Letting $u = \frac{1}{\beta^2 n}X^Ty$. (In the first step and in the last step, we tweak the coefficient of our first term assuming that we simply see according adjustments in our regularization term λ.)

$$
\begin{aligned}
\text{minimize}_w \frac{1}{n}\|Xw - y\|^2 + \lambda\|w\|^2 &\approx \text{minimize}_w \frac{1}{n^2}\|Xw - y\|^2 + \lambda\|w\|^2 \\
&\approx \text{minimize}_w \|\frac{1}{n}X^TXw - \frac{1}{n}X^Ty\|^2 + \lambda\|w\|^2 \\
&\approx \text{minimize}_w \|\beta^2 w - \frac{1}{n}X^Ty\|^2 + \lambda\|w\|^2 \\
&\approx \text{minimize}_w \beta^4\|w - \frac{1}{\beta^2 n}X^Ty\|^2 + \lambda\|w\|^2 \\
&\approx \text{minimize}_w \beta^2\|w - u\|^2 + \lambda\|w\|^2
\end{aligned}
$$

As in the scenario for least squares, we will again assume that $X \sim \mathcal{N}(0, \sigma^2)$ and $y = \beta^T x + e$, where $e = [\epsilon_1 \cdots \epsilon_n]^T$. We have that the solution to the above is

$$\hat{w} = (\frac{1}{1 + \lambda/\beta^2})u$$

2.2 Variance of Ridge Regression

We take our derivation for the variance of least squares, and get the following. From the above, since $\frac{1}{n}X^TX \approx \beta^2 I$, then $X^TX \approx n\beta^2 I$.

$$
\begin{aligned}
\mathbb{E}[\hat{y} - y] &= \mathbb{E}[x^T(X^TX + \lambda I)^{-1}X^Tee^TX(X^TX + \lambda I)^{-1}x] \\
&= \sigma^2\mathbb{E}[x^T(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}x] \\
&\approx n\beta^2\sigma^2\mathbb{E}[x^T(X^TX + \lambda I)^{-1}(X^TX + \lambda I)^{-1}x] \\
&\approx \frac{n\beta^2}{(n\beta^2 + \lambda I)^2}\sigma^2\mathbb{E}[x^Tx] \\
&\approx \frac{n\beta^2}{(n\beta^2 + \lambda I)^2}\sigma^2 d\beta^2 \\
&\approx \frac{n^2\beta^4}{(n\beta^2 + \lambda I)^2}\frac{\sigma^2 d}{n} \\
&\approx (\frac{n\beta^2}{n\beta^2 + \lambda I})^2\frac{\sigma^2 d}{n}
\end{aligned}
$$

We find that we can have small variance even when we have many more features than samples $d > n$, by adjusting λ accordingly.

3 Lasso

Lasso stands for "Least Absolute Shrinkage and Selection Operator" and often offers sparser solutions. We now consider a new objective function, where the penalty is an l1-norm.

$$
\sum_{i=1}^{n}(w^Tx_i - y_i)^2 + \lambda\|w\|_1
$$

With lasso, we make the following weight update:

$$
w_{k+1} = \text{shrink}(w_k - \alpha\nabla f(w_k))
$$

where we have the following definition of shrink.

$$
\text{shrink}(v)_i = \begin{cases} v_i - \alpha\lambda & v_i > \alpha\lambda \\ 0 & -\alpha\lambda < v_i < \alpha\lambda \\ v_i + \alpha\lambda & v_i < -\alpha\lambda \end{cases}
$$