Ridge Regression Example
Due when pigs fly

What better use of ridge regression than with ridge-back dragons?

Our trusty lab assistant has painstakingly measured the tail-length and ear-length of a set of local dragons, as well as the temperature of their fiery breath. The first two are measured in meters. The last is measured in "forgefires," the standard unit of measure for really hot things.

The data set looks like

$$X = \begin{bmatrix} 2 & 1 \\ 3 & 2 \\ 1 & 2 \\ 1 & 3 \\ 3 & 2 \\ 1 & 3 \end{bmatrix} \qquad Y = \begin{bmatrix} 10 \\ 5 \\ 7 \\ 4 \\ 11 \\ 6 \end{bmatrix} .$$

The first column of X are the tail-lengths for each dragon. The second column are the ear-lengths for the corresponding dragons. The elements of Y are the corresponding breadth temperatures. Because the temperature is the y value and the others the x features, we are trying to predict the temperature of a dragon's from the lengths of its tail and ear.

Doing linear regression, the final function will have the form

$$f(x) = w^\top x$$

where x is a column vector of the feature values for a dragon (2 elements in this case) and w is what we want to find out: the coefficients of the linear function.

Because we think that there may be a constant offset (thus the function should be $w^\top x + b$ for some constant b) and we want to also estimate this constant, we will incorporate the constant into the vector w by making w one element longer and adding a corresponding constant 1 to the data (this extra w element will always be multiplied by this constant 1 and therefore serve the same role as b in the equation above).

Thus we get

$$X = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 3 & 2 \\ 1 & 1 & 3 \end{bmatrix} \qquad Y = \begin{bmatrix} 10 \\ 5 \\ 7 \\ 4 \\ 11 \\ 6 \end{bmatrix} .$$

We will take the first 4 data points (dragons) to be our training set and the last 2 to be our validation set.

We will select three value of λ to try: $\{1, 10, 100\}$.

Let's try $\lambda = 1$ first:

We need to calculate

$$w = (X^\top X + \lambda \tilde{I})^{-1}(X^\top Y)$$

where

$$\tilde{I} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} . \qquad \text{(almost identity matrix, but upper-left element is zero instead of one)}$$

The calculations look like the following (remember, we are only using the first 4 data points).

$$X^\top X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 1 & 3 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 7 & 8 \\ 7 & 15 & 13 \\ 8 & 13 & 18 \end{bmatrix}$$

$$X^\top Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 1 \\ 1 & 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} 10 \\ 5 \\ 7 \\ 4 \end{bmatrix} = \begin{bmatrix} 26 \\ 46 \\ 46 \end{bmatrix}$$

$$X^\top X + \lambda \tilde{I} = \begin{bmatrix} 4 & 7 & 8 \\ 7 & 16 & 13 \\ 8 & 13 & 19 \end{bmatrix}$$

$$(X^\top X + \lambda \tilde{I})^{-1} = \begin{bmatrix} 3.29 & -0.707 & -0.902 \\ -0.707 & 0.293 & 0.098 \\ 0.902 & 0.098 & 0.366 \end{bmatrix}$$

$$w = (X^\top X + \lambda \tilde{I})^{-1}(X^\top Y) = \begin{bmatrix} 11.561 \\ -0.439 \\ -2.147 \end{bmatrix}$$

So the offset is 11.561 (that is, if a dragon has zero-length tail and ears, we'd predict this temperature). The two slopes are negative, which implies that as the tail and ears grow larger, the temperature drops.

We now need to find out how well this fits the validation data. We take the average squared error on the (two) validation points:

First Validation Point:

$$f\left(\begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}\right) = \begin{bmatrix} 1 & 3 & 2 \end{bmatrix} \begin{bmatrix} 11.561 \\ -0.439 \\ -2.147 \end{bmatrix} = 5.95 \ .$$

True answer should be 11 (from validation data). Therefore, squared error is 25.49.

Second Validation Point:

$$f\left(\begin{bmatrix} 1 \\ 1 \\ 3 \end{bmatrix}\right) = \begin{bmatrix} 1 & 1 & 3 \end{bmatrix} \begin{bmatrix} 11.561 \\ -0.439 \\ -2.147 \end{bmatrix} = 4.68 \ .$$

True answer should be 6 (from validation data). Therefore, squared error is 1.73

The average of 25.49 and 1.73 is $\boxed{13.61}$.

We now repeat for $\lambda = 10$:

$$X^\top X + \lambda \tilde{I} = \begin{bmatrix} 4 & 7 & 8 \\ 16 & 25 & 13 \\ 8 & 13 & 28 \end{bmatrix}$$

$$(X^\top X + \lambda \tilde{I})^{-1} = \begin{bmatrix} 0.873 & -0.151 & -0.180 \\ -0.151 & 0.079 & 0.007 \\ -0.180 & 0.007 & 0.084 \end{bmatrix}$$

$$w = (X^\top X + \lambda \tilde{I})^{-1}(X^\top Y) = \begin{bmatrix} 7.5 \\ 0 \\ -0.5 \end{bmatrix}$$

We now need to find out how well this fits the validation data. We take the average squared error on the (two) validation points:

First Validation Point:

$$f\left(\begin{bmatrix}1\\3\\2\end{bmatrix}\right) = \begin{bmatrix}1 & 3 & 2\end{bmatrix}\begin{bmatrix}7.5\\0\\-0.5\end{bmatrix} = 6.5\ .$$

True answer should be 11 (from validation data). Therefore, squared error is 20.25

Second Validation Point:

$$f\left(\begin{bmatrix}1\\1\\3\end{bmatrix}\right) = \begin{bmatrix}1 & 1 & 3\end{bmatrix}\begin{bmatrix}7.5\\0\\-0.5\end{bmatrix} = 6\ .$$

True answer should be 6 (from validation data). Therefore, squared error is 0

The average of 20.25 and 0 is $\boxed{10.13}$. This is better than the previous result, so we keep it (and remember $\lambda = 10$).

We now repeat for $\lambda = 100$:

$$X^\top X + \lambda\tilde{I} = \begin{bmatrix}4 & 7 & 8\\16 & 115 & 13\\8 & 13 & 118\end{bmatrix}$$

$$(X^\top X + \lambda\tilde{I})^{-1} = \begin{bmatrix}0.320 & -0.017 & -0.020\\-0.017 & 0.010 & 0\\-0.020 & 0 & 0.010\end{bmatrix}$$

$$w = (X^\top X + \lambda\tilde{I})^{-1}(X^\top Y) = \begin{bmatrix}6.61\\0.004\\-0.059\end{bmatrix}$$

We now need to find out how well this fits the validation data. We take the average squared error on the (two) validation points:

First Validation Point:

$$f\left(\begin{bmatrix}1\\3\\2\end{bmatrix}\right) = \begin{bmatrix}1 & 3 & 2\end{bmatrix}\begin{bmatrix}6.61\\0.004\\-0.059\end{bmatrix} = 6.51\ .$$

True answer should be 11 (from validation data). Therefore, squared error is 20.20

Second Validation Point:

$$f\left(\begin{bmatrix}1\\1\\3\end{bmatrix}\right) = \begin{bmatrix}1 & 1 & 3\end{bmatrix}\begin{bmatrix}6.61\\0.004\\-0.059\end{bmatrix} = 6.44\ .$$

True answer should be 6 (from validation data). Therefore, squared error is 0.19

The average of 20.20 and 0.19 is $\boxed{10.20}$. This is (slightly) worse than the previous best, so we do not keep it.

So our "answer" is the weight vector $w = \begin{bmatrix} 7.5 \\ 0 \\ -0.5 \end{bmatrix}$. If we wanted to separate this back out into an offset and a (normal) weight vector, we would have $b = 7.5$, $w = \begin{bmatrix} 0 \\ -0.5 \end{bmatrix}$.

Later (during testing or use "in the real world") if we run across a dragon with a tail length of 2 and an ear length of 4, we can plug this into our function as

$$\begin{aligned} f\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix}\right) &= \begin{bmatrix} 2 \\ 4 \end{bmatrix}^\top \begin{bmatrix} 0 \\ -0.5 \end{bmatrix} + 7.5 \\ &= \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}^\top \begin{bmatrix} 7.5 \\ 0 \\ -0.5 \end{bmatrix} \\ &= 5.5 \end{aligned}$$

and predict that the temperature of the flames out of said dragon's mouth would be 5.5 forgefires.