7.3

Intro

In this exercise we will derive the closed form solution for ridge regression. We will assume that $\bar{x} = 0$, which is a very reasonable assumption since most people have the habit of standardize their data before training a regressor. The problem in itself is very simple to understand: we have a cost function and our goal is to find its global minimum. Since this function is convex, find the global optimal is the same as taking the derivative of the cost function, equal it to zero and solve for (w_0, w). To take the derivatives, we could do it from scratch or we could use our knowledge of the ordinary least square solution and some matrix multiplication properties. We will derive the solution based on the latter approach.

Solution

First, let's rewrite our original cost function $J(w_0, w) = (y - Xw - w_0 1)^T(y - Xw - w_0 1) + \lambda w^T w$ in a more conveninet form. The new form is

$$
\begin{aligned}
&J(w_0, w) = \frac{1}{2}\left(y - \begin{bmatrix} | & | \\ 1 & X \\ | & | \end{bmatrix}\begin{bmatrix} w_0 \\ w \end{bmatrix}\right)^T \left(y - \begin{bmatrix} | & | \\ 1 & X \\ | & | \end{bmatrix}\begin{bmatrix} w_0 \\ w \end{bmatrix}\right) + \frac{1}{2}\lambda w^T w = \\
&J(w_0, w) = \frac{1}{2}(y - \hat{X}\hat{w})^T(y - \hat{X}\hat{w}) + \frac{1}{2}\lambda w^T w
\end{aligned}
\tag{1}
$$

This new form just divided the original expression by 2 and embedded the bias term (w_0) in the first expression into the vector weights creating the new matrix/vector $\hat{X}/\hat{w}$. Now let's take the matrix derivative of the two terms separetly:

$$
\begin{aligned}
&\nabla_{\hat{w}} \frac{1}{2}(y - \hat{X}\hat{w})^T(y - \hat{X}\hat{w}) = \hat{X}^T\hat{X}\hat{w} - \hat{X}^T y \\
&\nabla_{\hat{w}} \frac{1}{2}\lambda w^T w = \lambda \begin{bmatrix} 0 \\ w \end{bmatrix}
\end{aligned}
\tag{2}
$$

Note that for computing the gradient of the first expression, we just used the same gradient of ordinary least squares (ols). The gradient is valid because our first term has the exactly same form as the ols cost function. In the second expression, is worth noticing the first term is 0 because the penalty term is not a function of the bias w_0. Putting together the two parts, our total gradient becomes:

$$
\begin{aligned}
&\nabla_{\hat{w}} J(\hat{w}) = \hat{X}^T\hat{X}\hat{w} - \hat{X}^T y + \lambda \begin{bmatrix} 0 \\ w \end{bmatrix} = \\
&\begin{bmatrix} — & 1 & — \\ — & X^T & — \end{bmatrix}\begin{bmatrix} | & | \\ 1 & X \\ | & | \end{bmatrix}\begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} — & 1 & — \\ — & X^T & — \end{bmatrix} y + \lambda \begin{bmatrix} 0 \\ w \end{bmatrix} = \begin{bmatrix} | \\ 0 \\ | \end{bmatrix}
\end{aligned}
\tag{3}
$$

Now we have to solve the equation for w. Let's begin with the first row. In the equations below $x_j^{(i)}$ refers to the i-th feature of instance j:

$$\begin{aligned}
&\begin{bmatrix} — & 1 & — \end{bmatrix} \begin{bmatrix} | & | \\ 1 & X \\ | & | \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} — & 1 & — \end{bmatrix} y = 0 \\
&N w_0 + \sum_i \sum_j x_j^{(i)} w_i = \sum y_i \\
&N w_0 + N \sum_i \hat{x}^{(i)} w_i = N \hat{y} \\
&w_0 = \hat{y}
\end{aligned} \tag{4}$$

In the last step above, we used the fact the the mean of every feature $x^{(i)}$ is equal to 0, according with our initial hypothesis.

As the last step, let's solve the remaining gradient equations all at once:

$$\begin{aligned}
&\begin{bmatrix} — & X^T & — \end{bmatrix} \begin{bmatrix} | & | \\ 1 & X \\ | & | \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - \begin{bmatrix} — & X^T & — \end{bmatrix} y + \lambda \begin{bmatrix} w \end{bmatrix} = \begin{bmatrix} | \\ 0 \\ | \end{bmatrix} = \\
&\begin{bmatrix} | & | \\ 0 & X^T X \\ | & | \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} - X^T y + \lambda w = \begin{bmatrix} | \\ 0 \\ | \end{bmatrix} = \\
&\lambda w + X^T X w = X^T y \\
&w = (\lambda I + X^T X)^{-1} X^T y
\end{aligned} \tag{5}$$

Notice that in the equations above we made of the following property:

$$\hat{x} = 0 \implies \begin{bmatrix} — & X^T & — \end{bmatrix} \begin{bmatrix} | \\ 1 \\ | \end{bmatrix} = 0$$

Conclusion

In this exercise we found the global optimal of the cost function for ridge regression. We saw how the zero mean assumption $\hat{x} = 0$ and the use of our knowledge of the solution of ordinary least squares was very useful and simplified our calculations. Also, we can notice that the only goal of the bias $w_0 = \hat{y}$ is to subtract the mean from y. Therefore, if we subtract the mean from the output before starting the ridge regression, there will be no need to introduce the bias term.