Exercise sheet: Linear regression

Solutions prepared by Magnus Ross and Mauricio A Álvarez

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time to solve compared to an exercise with (***), which is a more complex exercise.

1. (*) Given the two vectors,

$$\mathbf{x} = \begin{bmatrix} 1.3 \\ -2.0 \\ 4.1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 0.4 \\ -0.8 \\ -1.1 \end{bmatrix}.$$

compute their inner product and their outer product.

Answer:
From the lecture slides, we know that the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ of dimension m is given by

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{m} x_i y_i.$$

For the vectors given in the question then, we get that

$$\begin{aligned} \mathbf{x}^\top \mathbf{y} &= \begin{bmatrix} 1.3 & -2.0 & 4.1 \end{bmatrix} \begin{bmatrix} 0.4 \\ -0.8 \\ -1.1 \end{bmatrix} \\ &= (1.3 \times 0.4) + (-2.0 \times -0.8) + (4.1 \times -1.1) \\ &= -2.39 \end{aligned}$$

We can also think of an inner product as a matrix product between a $1 \times m$ matrix and a $m \times 1$ matrix. From the lecture slides, the outer product of two vectors $\mathbf{x}$ and $\mathbf{y}$ of dimension p is given by

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_p \\ x_2 y_1 & \cdots & x_2 y_p \\ \vdots & \vdots & \vdots \\ x_m y_1 & \cdots & x_m y_p \end{bmatrix}$$

Note that in the case of the outer product, the dimension of the two vectors need not be the same, whereas in the case of the inner product they must be. For the vectors in the question we get,

$$\begin{aligned} \mathbf{x}\mathbf{y}^\top &= \begin{bmatrix} 1.3 \\ -2.0 \\ 4.1 \end{bmatrix} \begin{bmatrix} 0.4 & -0.8 & -1.1 \end{bmatrix} \\ &= \begin{bmatrix} 1.3 \times 0.4 & 1.3 \times -0.8 & 1.3 \times -1.1 \\ -2.0 \times 0.4 & -2.0 \times -0.8 & -2.0 \times -1.1 \\ 4.1 \times 0.4 & 4.1 \times -0.8 & 4.1 \times -1.1 \end{bmatrix} \\ &= \begin{bmatrix} 0.52 & -1.04 & -1.43 \\ -0.8 & 1.6 & 2.2 \\ 1.64 & -3.28 & -4.5 \end{bmatrix} \end{aligned}$$

We can also think of an outer product as a matrix product between a $m \times 1$ matrix and a $1 \times p$ matrix. The following code snippet computes both the inner and outer product using NumPy:

```
import numpy as np

x = np.array([1.3, -2.0, 4.1])
y = np.array([0.4, -0.8, -1.1])

print(f"Inner product: {np.dot(x, y):.2f}")
print(f"Outer product: {np.outer(x, y)}")
```

2. (**) Let us define a matrix $\mathbf{W}$ of dimensions $n \times m$, a vector $\mathbf{x}$ of dimensions $m \times 1$ and a vector $\mathbf{y}$ of dimensions $n \times 1$. Write the following expression in matrix form

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j + \sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j}.$$

[HINT: if necessary define a vector of ones $\mathbf{1}_p = [1 \cdots 1]^\top$ of dimensions $p \times 1$, where p can be any number].

Answer:

For the first term, the sum $\sum_{j=1}^{m} w_{i,j}x_j$ can be written as $\mathbf{Wx}$. To obtain $\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j$, we premultiply $\mathbf{Wx}$ by $\mathbf{1}_n^\top$ leading to $\mathbf{1}_n^\top\mathbf{Wx}$. For the second term, the sum $\sum_{i=1}^{n} y_i w_{i,j}$ can be expressed as $\mathbf{y}^\top\mathbf{W}$. To obtain $\sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j}$, we postmultiply by $\mathbf{1}_m$ leading to $\mathbf{y}^\top\mathbf{W}\mathbf{1}_m$. We can finally write

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j + \sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j} = \mathbf{1}_n^\top\mathbf{Wx} + \mathbf{y}^\top\mathbf{W}\mathbf{1}_m.$$

3. (***) Show that using the ML criterion, the optimal value for σ_*^2 is given as in slide 40 of Lecture 4, this is,

$$\sigma_*^2 = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w}_*)^\top(\mathbf{y} - \mathbf{X}\mathbf{w}_*).$$

Answer:

In our model, we assume that $\sigma \neq 0$. Based on the lecture notes, we know that that the log likelihood function is given by,

$$LL\left(\mathbf{w}, \sigma^2\right) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}),$$

which we need to maximise with respect to σ_*^2 As we have seen, we can find the optimal $\mathbf{w}$ that maximises $LL\left(\mathbf{w}, \sigma^2\right)$ by taking the gradient $\frac{dLL\left(\mathbf{w},\sigma^2\right)}{d\mathbf{w}}$, equating to zero. Similarly, we can find the optimal σ that maximises $LL\left(\mathbf{w}, \sigma^2\right)$ by taking the gradient $\frac{dLL\left(\mathbf{w},\sigma^2\right)}{d\sigma}$, equating to zero and then solving the resulting equation for σ_* (if we get the optimal value for σ_*, we can easily get the optimal value for σ_*^2).

Taking the gradient of each term in $L\left(\mathbf{w}, \sigma^2\right)$ wrt σ, we get

$$\frac{d}{d\sigma}\left[-\frac{N}{2}\log(2\pi)\right] = 0$$

$$\frac{d}{d\sigma}\left[-\frac{N}{2}\log\sigma^2\right] = -\frac{N}{2}\frac{d}{d\sigma}\left[\log\sigma^2\right] = -\frac{N}{2}\frac{d}{d\sigma}\left[2\log\sigma\right] = -\frac{N}{2}2\frac{d}{d\sigma}\left[\log\sigma\right] = -N\frac{d}{d\sigma}\left[\log\sigma\right] = -\frac{N}{\sigma}$$

$$\begin{aligned}
\frac{d}{d\sigma}\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\right] &= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{d}{d\sigma}\left[\frac{1}{2\sigma^2}\right] \\
&= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{2}\frac{d}{d\sigma}\left[\frac{1}{\sigma^2}\right] \\
&= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{2}\frac{d}{d\sigma}\left[\sigma^{-2}\right] \\
&= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{2}\left[(-2)\sigma^{-3}\right] \\
&= (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\left[\sigma^{-3}\right] \\
&= (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3}
\end{aligned}$$

Putting these terms together, we get

$$\frac{d}{d\sigma}LL\left(\mathbf{w}, \sigma^2\right) = 0 - \frac{N}{\sigma} + (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3}$$

Now, equating to zero and solving for σ^2, we get

$$\begin{aligned}
0 - \frac{N}{\sigma} + (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3} &= 0 \\
-\frac{N}{\sigma} + (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3} &= 0 \\
(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3} &= \frac{N}{\sigma} \\
(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) &= N\sigma^2 \quad \text{(We assume: } \quad \sigma \neq 0) \\
(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{N} &= \sigma^2 \\
\sigma^2 &= \frac{1}{N}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})
\end{aligned}$$

We already have $\mathbf{w}_*$, the optimal value for $\mathbf{w}$, so we just plug this in to obtain,

$$\sigma_*^2 = \frac{1}{N}\left(\mathbf{y}-\mathbf{X}\mathbf{w}_*\right)^\top\left(\mathbf{y}-\mathbf{X}\mathbf{w}_*\right).$$

Question 4 is in two parts, I'll answer the first part here:

4. (***) Start with the full expression for the cost function when using l_2 regularisation with 1st-order polynomial linear regression, $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_*)^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_*) + \lambda||\boldsymbol{w}||_2^2$ Write out the last term using an inner product and differentiate the whole expression to solve for $\mathbf{w}$.

   I'm going to make a change to the question, and use the *normalised* error, i.e. the *mean* squared error, as this is what is used in some ML packages, and I want to include it here, previously we've used the sum squared error. However not including it is also valid, but will lead to slightly different answers. The only difference is the inclusion of the constant N. I've also noticed that some authors add a 'half' to the penalty term: $\frac{\lambda}{2}||\boldsymbol{w}||_2^2$, so I'll use that too, but these changes don't really make a material difference.

   So the objective function is now,

$$E = \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_*)^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_*) + \frac{\lambda}{2}||\boldsymbol{w}||_2^2$$

   Differentiating each term (we went through in lectures how to differentiate the first term, but now we add the $\frac{1}{N}$). The last term we write as $\lambda||\boldsymbol{w}||_2^2 = \lambda\boldsymbol{w}^\top\boldsymbol{w}$ whose derivative is just $2\lambda\boldsymbol{w}$ (see eq. 131 in the matrix cookbook).

$$\frac{\partial E}{\partial \boldsymbol{w}} = \frac{2}{N}(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{w} - \frac{2}{N}\boldsymbol{X}^\top\boldsymbol{y} + 2\frac{\lambda}{2}\boldsymbol{w}$$

   Set equal to zero,divide by 2, and combine the two $\boldsymbol{w}$:

$$0 = \left(\frac{1}{N}(\boldsymbol{X}^\top\boldsymbol{X}) + \frac{\lambda}{2}\boldsymbol{I}\right)\boldsymbol{w} - \frac{1}{N}\boldsymbol{X}^\top\boldsymbol{y}$$

   Rearrange, and multiply through by N:

$$\left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda}{2}N\boldsymbol{I}\right)\boldsymbol{w} = \boldsymbol{X}^\top\boldsymbol{y}$$

   Multiply both sides by $\left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda}{2}N\boldsymbol{I}\right)^{-1}$ assuming it's invertible,

$$\boldsymbol{w}_* = \left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\lambda N}{2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

4. (*) You are given a dataset with the following instances, $(x_1, y_1) = (0.8, -1.2), (x_2, y_2) = (-0.3, -0.6)$, and $(x_3, y_3) = (0.1, 2.4)$. Find the optimal value $\mathbf{w}_*$ used in ridge regression with a regularisation parameter $\lambda = 0.1$.

Answer:
We have that the optimum value for $\mathbf{w}$ in ridge regression is given by,

$$\mathbf{w}_* = \left(\mathbf{X}^\top \mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y},$$

recall that $\mathbf{I}$ here is the identity matrix, and N is the number of data, in this case 3. We need to compute this equation for the dataset given in the question. We can write the dataset as,

$$\mathbf{X} = \begin{bmatrix} 1 & 0.8 \\ 1 & -0.3 \\ 1 & 0.1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} -1.2 \\ -0.6 \\ 2.4 \end{bmatrix}.$$

Where we add 1's to the first column to account for the intercept in the regression model. We need to compute the formula for this data. First let's compute some of the necessary terms, recalling that matrix multiplication is computed via

$$\begin{aligned} \mathbf{AB} &= \mathbf{C}, \\ C_{i,j} &= \sum_k A_{ik} B_{kj}, \end{aligned}$$

so,

$$\begin{aligned} \mathbf{X}^\top \mathbf{X} + \frac{\lambda N}{2} &= \left(\begin{bmatrix} 3 & 0.6 \\ 0.6 & 0.74 \end{bmatrix} + \frac{0.1 \times 3}{2} \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \\ &= \begin{bmatrix} 3.15 & 0.6 \\ 0.6 & 0.89 \end{bmatrix}, \\ \mathbf{X}^\top \mathbf{y} &= \begin{bmatrix} 0.6 \\ -0.54 \end{bmatrix}. \end{aligned}$$

Next we will need the formula to invert a 2×2 matrix, which is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Finally we can obtain $\mathbf{w}^*$,

$$\mathbf{w}^* = \begin{bmatrix} 0.364 & -0.246 \\ -0.246 & 1.289 \end{bmatrix} \begin{bmatrix} 0.6 \\ -0.54 \end{bmatrix} = \begin{bmatrix} 0.351 \\ -0.843 \end{bmatrix} \tag{1}$$

This question can also be solved using a maths package like NumPy (or even Matlab), the following snippet gives the solution using NumPy, the code uses `np.eye(p)` for representing $\mathbf{I}_p$ and `np.linalg.solve(A, b)` for solving the linear system $\mathbf{Ax} = \mathbf{b}$. See also the Lab Notebook for Week 4.

```python
import numpy as np

X = np.array([[1.0, 1.0, 1.0], [0.8, -0.3, 0.1]]).T
y = np.array([-1.2, -0.6, 2.4])

N = 3
l = 0.1

A = X.T @ X + ((N * l)/2) * np.eye(2)
b = X.T@y

# Solve Ax = b equivient to x = A^{-1} b
w_star = np.linalg.solve(A, b)
print(f"Optimum w: {w_star}")
```

5. (***) Consider a regression problem for which each observed output y_n has an associated weight factor $r_n > 0$, such that the mean of weighted squared errors is given as

$$E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} r_n(y_n - \mathbf{w}^\top \mathbf{x}_n)^2,$$

where $\mathbf{w} = [w_0, \ldots, w_D]^\top$ is the vector of parameters, and $\mathbf{x}_n \in \mathbb{R}^{D+1\times 1}$ with $x_{n,0} = 1$.

(a) Starting with the expression above, write the mean of weighted squared errors in matrix form. You should include each of the steps necessary to get the matrix form solution. [HINT: a diagonal matrix is a matrix that is zero everywhere except for the entries on its main diagonal. The weight factors $r_n > 0$ can be written as the elements of a diagonal matrix $\mathbf{R}$ of size $N \times N$].
Answer:
We start by writing the sum as

$$E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} r_n(y_n - \mathbf{w}^\top \mathbf{x}_n)^2 = \frac{1}{N}\sum_{n=1}^{N} r_n y_n^2 - \frac{2}{N}\sum_{n=1}^{N} r_n y_n \mathbf{w}^\top \mathbf{x}_n + \frac{1}{N}\sum_{n=1}^{N} r_n \left(\mathbf{w}^\top \mathbf{x}_n\right)^2.$$

Using the HINT given above, each term of the sum can be expressed as

$$\sum_{n=1}^{N} r_n y_n^2 = \mathbf{y}^\top \mathbf{R}\mathbf{y}$$

$$-2\sum_{n=1}^{N} r_n y_n \mathbf{w}^\top \mathbf{x}_n = -2\mathbf{w}^\top \sum_{n=1}^{N} r_n y_n \mathbf{x}_n = -2\mathbf{w}^\top \mathbf{X}^\top \mathbf{R}\mathbf{y}$$

$$\sum_{n=1}^{N} r_n \left(\mathbf{w}^\top \mathbf{x}_n\right)^2 = \sum_{n=1}^{N} r_n \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} = \mathbf{w}^\top \left(\sum_{n=1}^{N} r_n \mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{R}\mathbf{X}\mathbf{w},$$

where $\mathbf{y} = [y_1, \ldots, y_N]^\top$ and $\mathbf{X}$ is a *design matrix*. Putting these terms together in the expression for the mean of weigthed squared errors, we get

$$\begin{aligned} E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} r_n(y_n - \mathbf{w}^\top \mathbf{x}_n)^2 &= \frac{1}{N}\left[\mathbf{y}^\top\mathbf{R}\mathbf{y} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}\right] \\ &= \frac{1}{N}\left[\mathbf{y}^\top\mathbf{R}\mathbf{y} - \mathbf{y}^\top\mathbf{R}\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}\right] \\ &= \frac{1}{N}\left[\mathbf{y}^\top\mathbf{R}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) - \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)\right] \\ &= \frac{1}{N}\left[\left(\mathbf{y}^\top\mathbf{R} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\right)\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)\right] \\ &= \frac{1}{N}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^\top\mathbf{R}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right). \end{aligned}$$

(b) Find the optimal value of $\mathbf{w}$, $\mathbf{w}_*$, that minimises the mean of weighted squared errors. The solution should be in matrix form. Use matrix derivatives.
Answer:
We start with the mean of weighted squared errors in matrix form as

$$E(\mathbf{w}) = \frac{1}{N}\left(\mathbf{y}^\top\mathbf{R}\mathbf{y} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}\right).$$

The two main results that we need are

$$\frac{d\mathbf{w}^\top\mathbf{a}}{d\mathbf{w}} = \mathbf{a}, \qquad \frac{d\mathbf{w}^\top\mathbf{A}\mathbf{w}}{d\mathbf{w}} = 2\mathbf{A}\mathbf{w}.$$

The derivative of $E(\mathbf{w})$ wrt $\mathbf{w}$ is then given as

$$\frac{dE(\mathbf{w})}{d\mathbf{w}} = -\frac{2}{N}\mathbf{X}^\top\mathbf{R}\mathbf{y} + \frac{2}{N}\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}.$$

Making the expression above equal to zero, we get

$$\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{R}\mathbf{y}.$$

The optimal value $\mathbf{w}^*$ is then given as

$$\mathbf{w}^* = \left(\mathbf{X}^\top\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{R}\mathbf{y}.$$

6. (*) A dataset is used to train a linear regression model with polynomial basis functions $\{\phi_i(x) = x^i\}_{i=1}^{M}$, where $M = 4$. Assume that the weight vector after training is equal to $\mathbf{w}_* = [0.5, -0.8, 1.2, 1.3, -0.3]^\top$. What would be the predicted value for this linear model when the input is $x = 2.5$?
Answer:

For basis function regression, we must first compute the value of each basis function at the input location, and as usual add the bias. In this case we have,

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 1 \\ 2.5^1 \\ 2.5^2 \\ 2.5^3 \\ 2.5^4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.5 \\ 6.25 \\ 15.625 \\ 39.0625 \end{bmatrix}$$

The prediction can then be computed by taking the inner product with the weight vector giving,

$$\mathbf{w}_*^\top \boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 0.5 & -0.8 & 1.2 & 1.3 & -0.3 \end{bmatrix} \begin{bmatrix} 1. \\ 2.5 \\ 6.25 \\ 15.625 \\ 39.0625 \end{bmatrix} = 14.594$$

Alternatively the following Python code computes the answer:

```python
import numpy as np

w_star = np.array([0.5, -0.8, 1.2, 1.3, -0.3])
phi = np.power(2.5 * np.ones(5), np.arange(5))

print(f"Prediction: {np.dot(w_star, phi)}")
```

7. (***) Show that the optimal solution for $\mathbf{w}_*$ in ridge regression is given as in slide 63 of Lecture 4, this is,

$$\mathbf{w}_* = \left(\mathbf{X}^\top \mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Answer:

Based on the Lecture notes, in ridge regression, we consider the objective function as

$$h(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}$$

Using what we reviewed in the section on vector/matrix notation, it can be shown that this expression can be written in a vectorial form as

$$h(\mathbf{w}) = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}$$

The term $\frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})$ can be expressed as

$$\frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) = \frac{1}{N}\mathbf{y}^\top \mathbf{y} - \frac{1}{N}\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} - \frac{1}{N}\mathbf{y}^\top \mathbf{X}\mathbf{w} + \frac{1}{N}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$$

We can find the $\mathbf{w}$ that maximises $h(\mathbf{w})$ by taking the gradient $\frac{dh(\mathbf{w})}{d\mathbf{w}}$, equating to zero and solving for $\mathbf{w}$. Taking the gradient of each term in $h(\mathbf{w})$ wrt $\mathbf{w}$, we get

$$\begin{aligned}
\frac{d}{d\mathbf{w}}\left[\frac{1}{N}\mathbf{y}^\top\mathbf{y}\right] &= 0 \\
\frac{d}{d\mathbf{w}}\left[-\frac{1}{N}\mathbf{w}^\top\mathbf{X}^\top\mathbf{y}\right] &= -\frac{1}{N}\mathbf{X}^\top\mathbf{y} \\
\frac{d}{d\mathbf{w}}\left[-\frac{1}{N}\mathbf{y}^\top\mathbf{X}\mathbf{w}\right] &= -\frac{1}{N}\mathbf{X}^\top\mathbf{y} \\
\frac{d}{d\mathbf{w}}\left[\frac{1}{N}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right] &= \frac{2}{N}\mathbf{X}^\top\mathbf{X}\mathbf{w} \\
\frac{d}{d\mathbf{w}}\left[\frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}\right] &= \frac{\lambda}{2}2\mathbf{w} = \lambda\mathbf{w}
\end{aligned}$$

Putting these terms together, we get

$$\begin{aligned}
\frac{d}{d\mathbf{w}}h\left(\mathbf{w}\right) &= 0 - \frac{1}{N}\mathbf{X}^\top\mathbf{y} - \frac{1}{N}\mathbf{X}^\top\mathbf{y} + \frac{2}{N}\mathbf{X}^\top\mathbf{X}\mathbf{w} + \lambda\mathbf{w} \\
&= -\frac{2}{N}\mathbf{X}^\top\mathbf{y} + \frac{2}{N}\mathbf{X}^\top\mathbf{X}\mathbf{w} + \lambda\mathbf{w} \\
&= -\frac{2}{N}\mathbf{X}^\top\mathbf{y} + \left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w},
\end{aligned}$$

where $\mathbf{I}$ is an identity matrix of the same dimensions that $\mathbf{w}$. Now, equating to zero and solving for $\mathbf{w}$, we get

$$\begin{aligned}
-\frac{2}{N}\mathbf{X}^\top\mathbf{y} + \left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= 0 \\
\left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= \frac{2}{N}\mathbf{X}^\top\mathbf{y} \\
\frac{N}{2}\left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= \frac{N}{2}\frac{2}{N}\mathbf{X}^\top\mathbf{y} \\
\left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)\mathbf{w} &= \mathbf{X}^\top\mathbf{y} \\
\mathbf{w} &= \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}$$

Thus, the optimal solution $\mathbf{w}_*$ in ridge regression is

$$\mathbf{w}_* = \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$$