Exercise sheet: Linear regression

Solutions prepared by: Mr Chunchao Ma (PhD Candidate), Mauricio A Álvarez

1. 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

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

2. Show that using the ML criterion, the optimal value for σ_*^2 is given as in slide 40 of Lecture 3, 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 note, we already know that

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

As we have seen in the Lecture, 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 we can get the optimal value 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

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

From the lecture notes, we already know $\mathbf{w}_*$ is the optimal value for $\mathbf{w}$. We put $\mathbf{w}_*$ so we get

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

3. 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

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

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

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

4. Show that the optimal solution for $\mathbf{w}_*$ in ridge regression is given as in slide 63 of Lecture 3, 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

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

5. 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 first build the design matrix

$$\mathbf{X} = \begin{bmatrix}\mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \mathbf{x}_3^\top\end{bmatrix} = \begin{bmatrix}1 & 0.8 \\ 1 & -0.3 \\ 1 & 0.1\end{bmatrix}.$$

   The vector $\mathbf{y}$ has elements $\mathbf{y} = [-1.2, -0.6, 2.4]^\top$. Applying the expression above,

$$\begin{aligned}\mathbf{w}_* &= \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \left(\begin{bmatrix}1 & 1 & 1 \\ 0.8 & -0.3 & 0.1\end{bmatrix}\begin{bmatrix}1 & 0.8 \\ 1 & -0.3 \\ 1 & 0.1\end{bmatrix} + \frac{(0.1)(3)}{2}\begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1 & 1 \\ 0.8 & -0.3 & 0.1\end{bmatrix}\begin{bmatrix}-1.2 \\ -0.6 \\ 2.4\end{bmatrix} \\ &= \begin{bmatrix}0.35 \\ -0.84\end{bmatrix}\end{aligned}$$

Figure 1 shows a snippet of Python code that computes $\mathbf{w}_*$,

```
In [25]: import numpy as np
    ...:
    ...: lambdap = 0.1
    ...: X = np.array([[1, 0.8],[1, -0.3], [1, 0.1]])
    ...: N, D = np.shape(X)
    ...: y = np.array([[-1.2], [-0.6], [2.4]])
    ...: w = np.linalg.solve(X.T@X + (lambdap*N/2)*np.eye(D), X.T@y)
    ...: w
Out[25]:
array([[ 0.35113567],
       [-0.84346225]])
```

Figure 1: Snippet of Python code that computes $\mathbf{w}_*$

The Python code uses `np.eye(p)` for representing $\mathbf{I}_p$ and `np.linalg.solve(A, b)` for solving the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$. See also the Lab Notebook for Week 3.