Lund University
Faculty of Engineering (LTH)

FMAN45

Machine Learning

Assignment 1

Author:
Marcel Attar, 941127-2173

The report was handed in on: April 14, 2019

LUNDS UNIVERSITET
Lunds Tekniska Högskola

Task 1

We want to solve the equation

$$\text{minimize}_{w_i} \frac{1}{2} \left\| \mathbf{r_i} - \mathbf{x_i} w_i \right\|_2^2 + \lambda \left| w_i \right|, \lambda \geq 0 \tag{1}$$

By differentiating we get the closed-form

$$\frac{\mathrm{d}}{\mathrm{dw_i}} \left(\frac{1}{2} \left\| \mathbf{r_i} - \mathbf{x_i} w_i \right\|_2^2 + \lambda \left| w_i \right| \right) = 0$$

Given that $w_i \neq 0$ we get

$$\frac{\mathrm{d}}{\mathrm{dw_i}} \left(\frac{1}{2} \left\| \mathbf{r_i} - \mathbf{x_i} w_i \right\|_2^2 \right) + \lambda \frac{w_i}{|w_i|} = 0$$

Then

$$\frac{1}{2} 2 (\mathbf{r_i} - \mathbf{x_i} w_i)^T (-\mathbf{x_i}) + \lambda \frac{w_i}{|w_i|} = 0 \quad \Leftrightarrow$$

$$(\mathbf{x_i}^T \mathbf{x_i} + \frac{\lambda}{|w_i|}) w_i = \mathbf{r}^T \mathbf{x} \tag{2}$$

And since $\mathbf{x_i}^T \mathbf{x_i} \geq 0$ and $\frac{\lambda}{|w_i|} \geq 0$ we can take the absolute value of both sides, resulting in

$$\Rightarrow \quad (\mathbf{x_i}^T \mathbf{x_i} + \frac{\lambda}{|w_i|}) \left| w_i \right| = \left| \mathbf{r}^T \mathbf{x} \right| \quad \Leftrightarrow$$

$$|w_i| = \frac{|\mathbf{r_i}^T \mathbf{x_i}| - \lambda}{\mathbf{x_i}^T \mathbf{x_i}} \tag{3}$$

From equation 2 we get that

$$w_i = \frac{\mathbf{r_i}^T \mathbf{x_i}}{\mathbf{x_i}^T \mathbf{x_i} + \frac{\lambda}{|w_i|}}$$

If we now substitute $|w_i|$ with the value from equation 3 we get the first line in

$$w_i^{(j)} = \begin{cases} \frac{\mathbf{x_i}^T \mathbf{r_i}^{(j-1)}}{\mathbf{x_i}^T \mathbf{x_i} |\mathbf{x_i}^T \mathbf{r_i}^{(j-1)}|} (|\mathbf{x_i}^T \mathbf{r_i}^{(j-1)}| - \lambda), & |\mathbf{x_i}^T \mathbf{r_i}^{(j-1)}| > \lambda \\ 0, & |\mathbf{x_i}^T \mathbf{r_i}^{(j-1)}| \leq \lambda \end{cases} \tag{4}$$

Task 2

We want to show that

$$\hat{w}_i^{(2)} - \hat{w}_i^{(1)} = 0, \quad \forall i \tag{5}$$

given that $\mathbf{X^T X} = \mathbf{I_N}$. We get that

$$\mathbf{x_i}^T \mathbf{r_i}^{(j-1)} = \mathbf{x_i}^T (\mathbf{t} - \sum_{l \neq 1} \mathbf{x_l} \mathbf{\hat{w}_l^{(j-1)}}) = \mathbf{x_i}^T \mathbf{t} \tag{6}$$

since $\mathbf{X}$ is a orthonormal basis. If we now use this in equation 4 we get

$$w_i^{(j)} = \frac{\mathbf{x_i}^T \mathbf{r_i}}{\mathbf{x_i}^T \mathbf{x_i} \, |\mathbf{x_i}^T \mathbf{t}|} (|\mathbf{x_i}^T \mathbf{t}| - \lambda) \tag{7}$$

Since the right side is not dependant on j we have proven that 5 holds.

Task 3

We want to calculate

$$\lim_{\sigma \to 0} (\mathbb{E}(\hat{w}_i{}^{(1)} - w_i^*)) \tag{8}$$

given that $\mathbf{t} = \mathbf{X}\mathbf{w}^* + \mathbf{e}$ and $\mathbf{e} \sim N(\mathbf{0}_N, \sigma \mathbf{I}_N)$, N is a Gaussian distribution. Starting to solve the equation gives

$$\mathbb{E}(\hat{w}_i{}^{(1)} - w_i^*) = \mathbb{E}(\hat{w}_i{}^{(1)}) - \mathbb{E}(w_i^*) = \mathbb{E}(\hat{w}_i{}^{(1)}) - w_i^*$$

In the last step we used that w_i^* is a non-random variable. We now want to solve

$$\mathbb{E}(\hat{w}_i{}^{(1)}) = \mathbb{E}(\frac{\mathbf{r_i}^{(0)T} \mathbf{x_i}}{\mathbf{x_i}^T \mathbf{x_i} \, |\mathbf{x_i}^T \mathbf{t}^{(0)}|} (|\mathbf{x_i}^T \mathbf{t}^{(0)}| - \lambda)) \tag{9}$$

$$\mathbf{r}_i^{(j-1)} = \mathbf{t} - \sum_{l \neq i} x_l \hat{w}_l^{(j-1)} = \mathbf{X}\mathbf{w}^* + \mathbf{e} - \sum_{l \neq i} x_l \hat{w}_l^{(j-1)} \tag{10}$$

Using 10 in equation 9, the condition that $\mathbf{r_i}^{(0)T} \mathbf{x_i} > \lambda$ and that $\mathbf{X}$ is orthonormal gives us

$$\mathbb{E}(\hat{w}_i{}^{(1)}) = \mathbb{E}(\frac{\mathbf{r_i}^{(0)T} \mathbf{x_i} - \lambda}{\mathbf{x_i}^T \mathbf{x_i}}) = \mathbb{E}(\frac{\mathbf{x_i}^T (\mathbf{X}\mathbf{w}^* + \mathbf{e} - \sum_{l \neq i} \mathbf{x_l} \hat{w}_l^{(j-1)}) - \lambda}{\mathbf{x_i}^T \mathbf{x_i}}) \quad \Leftrightarrow$$

$$\mathbb{E}(\hat{w}_i{}^{(1)}) = \mathbb{E}(\mathbf{x_i}^T \mathbf{X}\mathbf{w}^* + \mathbf{x_i}^T \mathbf{e} - \lambda) = \mathbb{E}(\mathbf{x}_i^T \mathbf{w}^*) - \lambda = \mathbf{w}_i^* - \lambda$$

Now the same approach, solving equation 9 but using the condition $\mathbf{r_i}^{(0)T} \mathbf{x_i} < -\lambda$ gives us

$$\mathbb{E}(\hat{w}_i{}^{(1)}) = \mathbb{E}(\mathbf{r_i}^{(0)T} \mathbf{x_i} + \lambda) = \mathbb{E}(\mathbf{x_i}^T (\mathbf{X}\mathbf{w}^* + \mathbf{e} - \sum_{l \neq i} \mathbf{x_l} \hat{w}_l^{(j-1)}) + \lambda) = w_i^* + \lambda$$

And finally for the condition $|\mathbf{r_i}^{(0)T} \mathbf{x_i}| \leq \lambda$

$$\mathbb{E}(\hat{w}_i{}^{(1)}) = \mathbb{E}(0) = 0$$

Before we use these values of $\mathbb{E}(\hat{w}_i{}^{(1)})$ in equation 11 we need take the limit of the left hand side of the conditions.

$$\lim_{\sigma \to 0} \mathbf{r_i}^{(0)T} \mathbf{x_i} = \lim_{\sigma \to 0} \mathbf{x}_i^T (\mathbf{X}\mathbf{w}^* + \mathbf{e} - \sum_{l \neq i} \mathbf{x_l} \hat{w}_l^{(j-1)}) = \lim_{\sigma \to 0} (w_i^* + x_i^T \mathbf{e}) = w_i^*$$

Finally we get

$$\lim_{\sigma \to 0} (\mathbb{E}(\hat{w}_i{}^{(1)} - w_i^*)) = \begin{cases} -\lambda, & w_i^* > \lambda \\ -w_i^*, & |w_i^*| \leq \lambda \\ \lambda, & w_i^* < -\lambda \end{cases} \tag{11}$$

This estimation bias that is generated using LASSO, Least Absolute Shrinkage and Selection Operator. There are two parts: (1) shrinkage and (2) variable selection. (1) We want to shrink our λ so that the bias get smaller and (2) LASSO wants to select a few w_i's and the rest get value 0.

Task 4

Using the provided skeleton code, equation 4 was implemented and the following graphs were produced

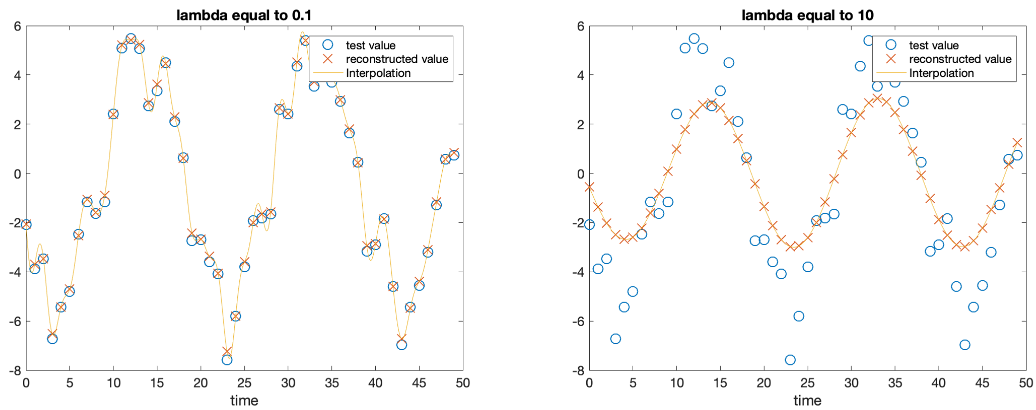

Figure 1: Graphs showing the reconstructed values for two different λ, $\lambda = 0.1$ to the left and $\lambda = 10$ to the right. Smaller λ leads to overfitting (left graph), bigger λ results in underfitting (right graph).

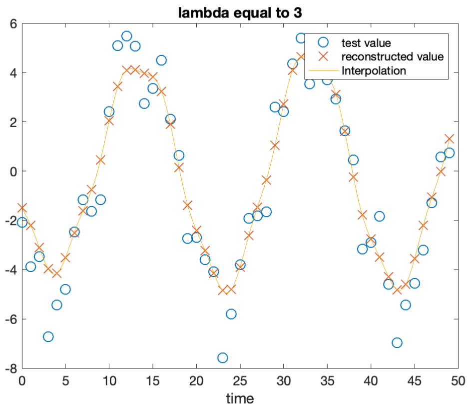

Figure 2: By choosing $\lambda = 3$ the reconstructed points follow the test values without overfitting.

Choosing a big λ results in undefitting. By looking at equation 1, which is the main LASSO equation, it is clear that a big λ will penalize large values of w_i, therefore promoting small w_i's which leads to underfitting. The opposite is the case for a small λ. The most suitable λ value that I could find, for this case, was 3. The corresponding reconstruction plot can be seen in figure 2.

Counting the number of $\hat{w}_i$ that are equal to zero for the three λ values give

λ-value	**Number of non-zero coordinates**
0.1	233
10	6
3	13

The actual number of non-zero coordinates needed to model the data is 4. Therefore, it seems like $\lambda = 10$ is the optimal value, out of the three λ values, with respect to this.

Task 5

The LASSO algorithm and K-fold cross validation schema was implemented using the provided skeleton code. Using that the following two plots were produced

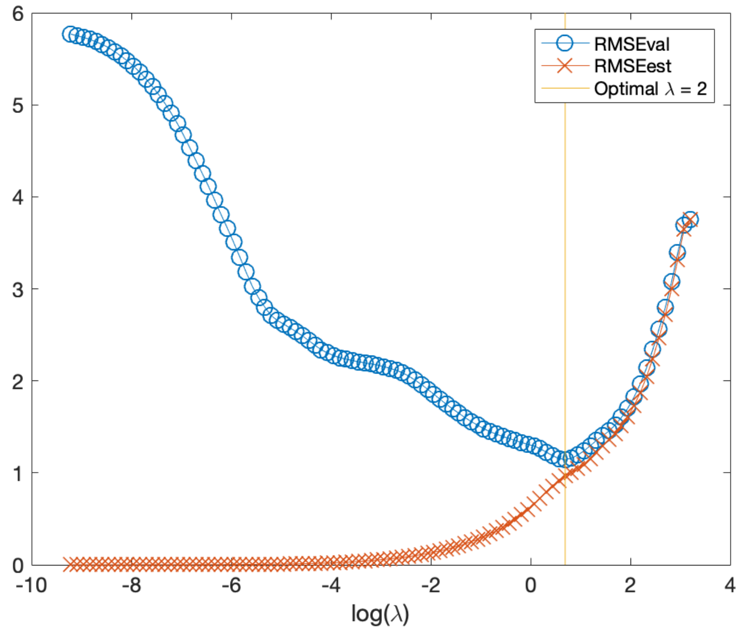

Figure 3: The graph shows the root mean square error for both the validation data and the estimate data.

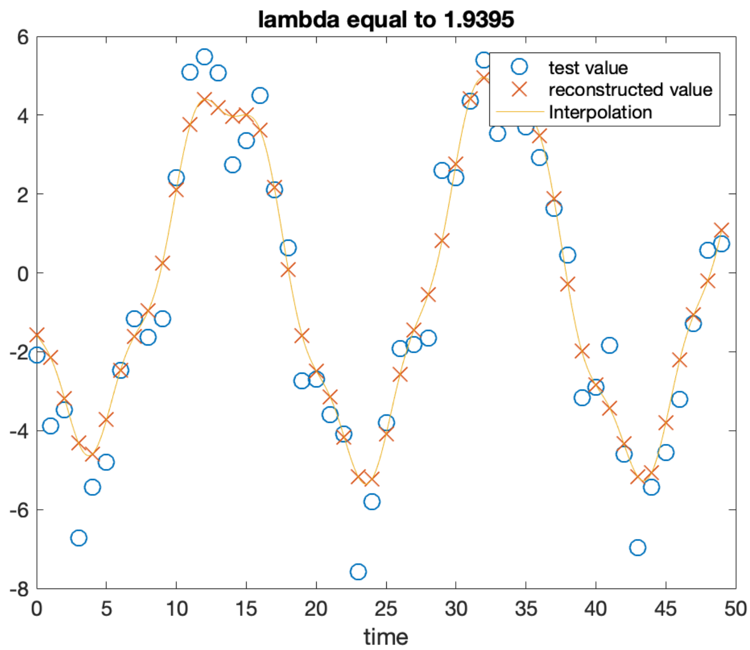

Figure 4: The plot given by choosing the optimal λ given by the K-fold cross validation.

As can be seen in figure 3 the optimal λ value is approximately 2, this is the point where the root mean square error for the validation data is minimized.

This value can be compared with the user generated λ, which was 3. The K-fold cross validation scheme therefore generated a smaller λ, accepting more fitting (potentially overfitting). By looking at figure 4 the overfitting tendencies can be observed since the interpolation is somewhat jagged.

Task 6

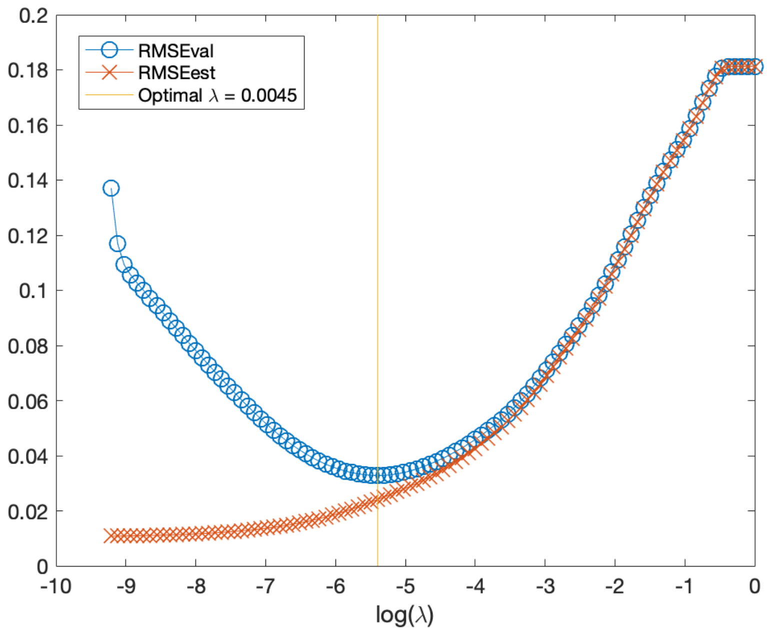

Figure 5: The graph shows the root mean square error for both the validation data and the estimate data.

The shapes of the two RMSE lines in figure 5 are similar to figure 3 but with smaller values for both λ and the errors. When λ becomes large enough the RMSE is constant, as can be seen in the top right corner of the graph. This is due to the condition $|\mathbf{x_i}^T \mathbf{r_i}^{(j-1)}| \leq \lambda$ in equation 4 which will be true for all coordinates when λ becomes large, resulting in $w_i^{(j)} = 0$ which leads to the error being constant.

Task 7

Listen to attached sound file. The background noise gets removed to some extent when using the provided lasso_denoise algorithm and using the the optimal λ value that was calculated in task 6.