FMAN45 Machine Learning

Assignment 1

(Van Duy Dang - va7200da-s)

April 23, 2023

1 Penalized regression via the LASSO

Task 1:

The objective is to

$$\underset{w_i}{\text{minimize}}\ \frac{1}{2}\|r_i - x_i w_i\|_2^2 + \lambda|w_i| \tag{1}$$

We find w_i by differentiating (1) and solve the equation

$$\begin{aligned}
&\frac{d}{dw_i}(\frac{1}{2}\|r_i - x_i w_i\|_2^2 + \lambda|w_i|) = 0 \\
\iff\ &\frac{1}{2}(2(r_i - x_i w_i)(-x_i)) + \lambda\frac{w_i}{|w_i|} = 0 \\
\iff\ &x_i^T r_i = w_i(x_i^T x_i + \frac{\lambda}{|w_i|}) \qquad (2)
\end{aligned}$$

Then we take the absolute value of both sides.

$$|x_i^T r_i| = |w_i||(x_i^T x_i + \frac{\lambda}{|w_i|})| \tag{3}$$

Because $x_i^T x_i$ and λ are positive so from (3) we have:

$$\begin{aligned}
&|x_i^T r_i| = |w_i|(x_i^T x_i + \frac{\lambda}{|w_i|}) \\
\iff\ &|w_i| = \frac{|x_i^T r_i| - \lambda}{x_i^T x_i} \qquad (4)
\end{aligned}$$

We use (2) to find $\text{sgn}(w_i)$. Since $x_i^T x_i$ and λ are positive

$$sgn(w_i) = sgn(x_i^T r_i) = \frac{x_i^T r_i}{|x_i^T r_i|} \tag{5}$$

Now w_i can be calculated to verify task 1 by multiplying (4) and (5)

$$w_i = \frac{x_i^T r_i}{x_i^T x_i |x_i^T r_i|}(|x_i^T r_i| - \lambda) \tag{6}$$

Task 2:

Since $x_i^T x_l = 0, \forall l \neq i$, we have

$$x_i^T r_i^{(j-1)} = x_i^T(t - \sum_{l<i}(x_l \hat{w}_l^{(j)}) - \sum_{l>i}(x_l \hat{w}_l^{(j-1)})) = x_i^T t$$

With $x_i^T x_i = 1$, we can rewrite (6) as

$$\hat{w}_i^{(j)} = x_i^T t - \lambda sgn(x_i^T t) \tag{7}$$

We can see that $\hat{w}_i^{(j)}$ does not depend on previous estimates. Therefore

$$\hat{w}_i^{(2)} - \hat{w}_i^{(1)} = x_i^T t - \lambda sgn(x_i^T t) - (x_i^T t - \lambda sgn(x_i^T t)) = 0$$

In addition, this is also correct in case of $|x_i^T r_i^{(j-1)}| \leq \lambda$ as $\hat{w}_i^{(j)} = 0$

Task 3:

We have $t = Xw^* + e$. The limit of $x_i^T t$ when $\sigma \to 0$ is:

$$\lim_{\sigma \to 0} x_i^T t = x_i^T X w^* = w^*$$

Considering the first case from the hint where $x_i^T r_i^{(j-1)} > \lambda$, we use (7) to calculate

$$\begin{aligned}
\lim_{\sigma \to 0} E(\hat{w}_i^{(1)} - w_i^*) &= \lim_{\sigma \to 0} E(x_i^T t - \lambda sgn(x_i^T t) - w_i^*) \\
&= \lim_{\sigma \to 0} E(w_i^* - \lambda sgn(w_i^*) - w_i^*) \\
&= \lim_{\sigma \to 0} E(-\lambda sgn(w_i^*)) = -\lambda
\end{aligned}$$

The second case is $x_i^T r_i^{(j-1)} < \lambda$. We use the same equation but $sgn(w_i^*) = -1$. Hence

$$\lim_{\sigma \to 0} E(\hat{w}_i^{(1)} - w_i^*) = \lim_{\sigma \to 0} E(-\lambda sgn(w_i^*)) = \lambda$$

The third case is $|x_i^T r_i^{(j-1)}| \leq \lambda$. From the requirement, we have $\hat{w}_i^{(1)} = 0$. The limit can be calculated as

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

In conclusion, we get all 3 cases that we need to show.

LASSO (Least Absolute Shrinkage and Selection Operator) is a regularization technique used in linear regression models to prevent overfitting by shrinking the estimated coefficients towards zero. When the penalty parameter λ is sufficiently large, the LASSO method can perform variable selection effectively by forcing many of the estimated coefficients to be exactly zero.

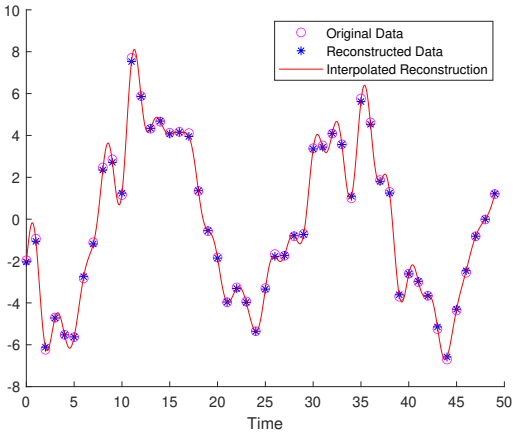
The limit of the expected difference between $\hat{w}_i^{(1)}$ and w_i^*, as described above, shows that the LASSO method shrinks $\hat{w}_i^{(1)}$ towards zero, with the amount of shrinkage depending on w_i^* and λ.

2 Hyperparameter-learning via K-fold cross-validation

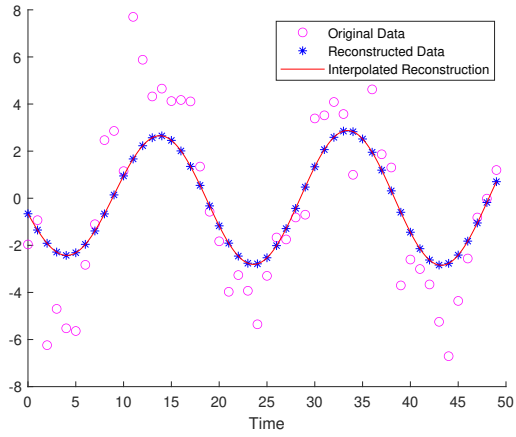
Task 4:

We can see that $\lambda = 0.1$ is small so the model is overfitting, it fits the original data. With $\lambda = 10$, it seems that the reconstructed data is shrunk too much and the model is underfitting. In my opinion, $\lambda = 1.5$ is a good value. The reconstruction data looks similar

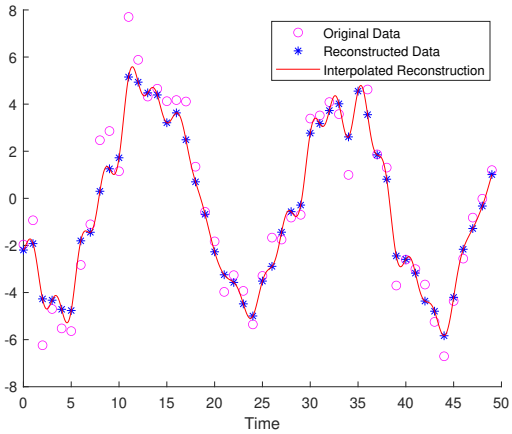
to the figure from the hint. The model does not fit the noise and it generalizes better than the previous cases.



(a) $\lambda = 0.1$



(b) $\lambda = 10$



(c) $\lambda = 1.5$

Figure 1: Reconstruction plot

From Table 1, we can see that the smaller λ the larger the number of non-zero coordinates. The best λ that I choose has 66 non-zero coordinates. In other words, we need more than 4 non-zero coordinates to find the good model.

λ	**Number of non-zero coordinates**
0.1	265
10	8
1.5	66

Table 1: Number of non-zero coordinates

Task 5:

In this task, we implement a K-fold cross-validation scheme for the LASSO estimator to find the optimal λ. The training data is randomly split in 10 folds and λ is a grid of 100

values.

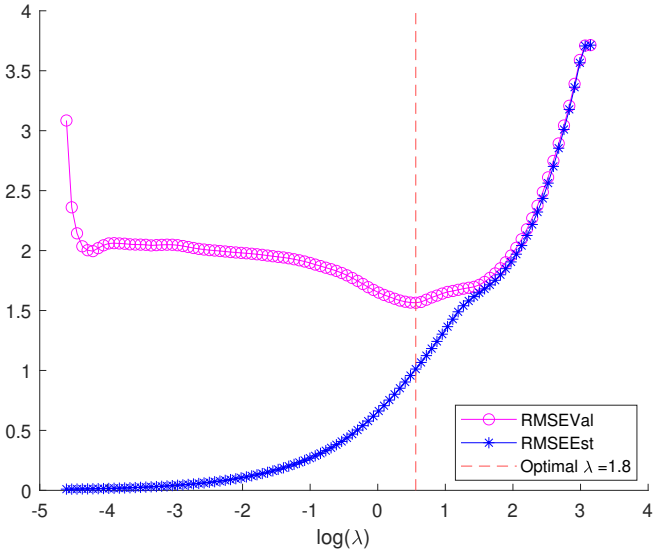


Figure 2: $RMSE_{val}, RMSE_{est}$ and the optimal λ

Here we can visualize the results we got from task 4. With small λ values, $RMSE_{val}$ is significantly lower than $RMSE_{est}$. It means that the model fits the training data well but the validation data. In other words, the model is overfitting. When λ is large, $RMSE_{val}$ and $RMSE_{est}$ are also large. It means that the model is underfitting and it does not perform well on both training data and validation data. We find the best model by finding $\lambda = 1.8$ that minimizes $RMSE_{val}$.

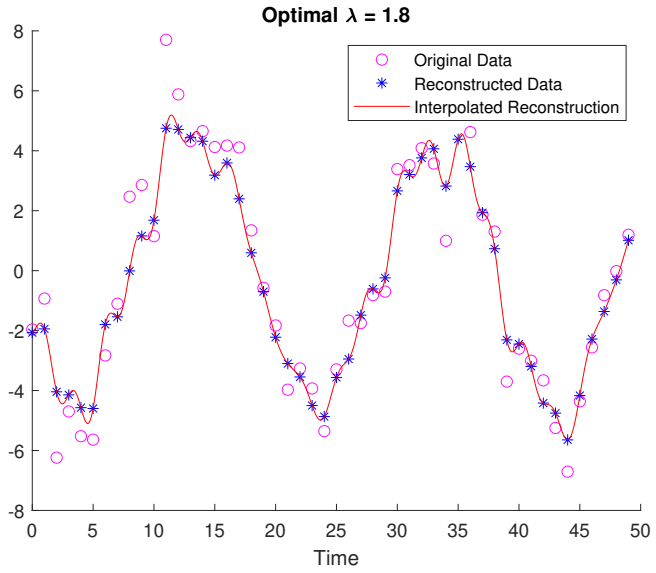


Figure 3: Reconstruction plot with $\lambda = 1.8$

The result looks similar to the final plot from task 4 and the plot from the hint.

3 Denoising of an audio excerpt

Task 6:

In this task, I choose the number of folds is 3 so that the training is faster and the data is divided evenly. The plot is similar to task 5 and we can find the optimal $\lambda = 0.0038$ that minimizes $RMSE_{val}$.

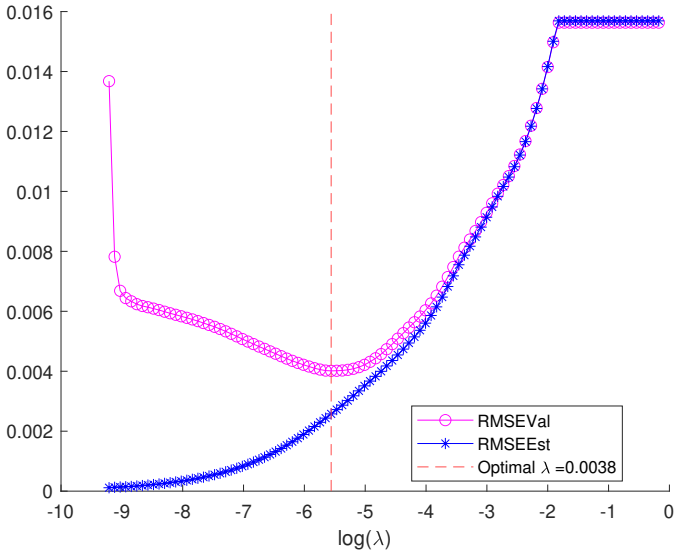


Figure 4: $RMSE_{val}, RMSE_{est}$ and the optimal $\lambda = 0.0038$

At the top right of the plot, it seems that both $RMSE_{val}$ and $RMSE_{est}$ are constant. This is because λ is too large and the model is underfitting. Increasing λ does not have any significant impact on the performance of the model.

Task 7:

After denoising the test data with $\lambda = 0.0038$, I feel that the background noise has been filtered out to a small extent. In my opinion, the optimal λ found from task 6 is not a good value to denoise. With $\lambda = 0.015$, all the noise seems to be filtered out.