Introduction to Machine Learning (67577)
Exercise V – Validation, Feature Selection and Regularization
Hadar Sharvit – 208287599

Theoretical Questions

Validation

1.

A. Bounding the generalization error using the *standard method*

From Hoeffding's inequality we know that $P[|\bar{X}-E[\bar{X}]|\geq\varepsilon]\leq 2e^{-2m\varepsilon^2}$, therefore, $P[|L_{S_{all}}(h)-L_D(h)|\geq\varepsilon]\leq 2e^{-2m\varepsilon^2}$. If we set $\varepsilon=\sqrt{\frac{\ln(2/\delta)}{2m}}$ we have $P\left[|L_{S_{all}}(h)-L_D(h)|\leq\sqrt{\frac{\ln(2/\delta)}{2m}}\right]\geq 1-\delta$. Inverting the signs $P\left[|L_{S_{all}}(h)-L_D(h)|\geq\sqrt{\frac{\ln(2/\delta)}{2m}}\right]\leq\delta$. Using Hoeffding's yet again while adding $|H_k|$ to the ln, we have $P\left[|L_{S_{all}}(h)-L_D(h)|\geq\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}\right]\leq\frac{\delta}{|H_k|}$

This is true for every h, therefore using Union bound over all h_i we get

$$P\left[\bigcup_{h_i\in H_k}\left(|L_{S_{all}}(h_i)-L_D(h_i)|\geq\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}\right)\right]\leq|H_k|\cdot\frac{\delta}{|H_k|}=\delta$$

So, inverting the probabilities yet again - for every h_i we have $P\left(|L_{S_{all}}(h_i)-L_D(h_i)|\leq\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}\right)\geq 1-\delta$. Let us concentrate in the inner section of the probability, and for h^*

$$L_{S_{all}}(h^*)-L_D(h^*)\leq\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}\rightarrow L_D(h^*)\leq L_{S_{all}}(h^*)+\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}$$

h^* is better than any h_i, so

$$\leq L_{S_{all}}(h_i)+\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}$$

Substituting $L_{S_{all}}(h_i)$ again:

$$\leq L_D(h_i)+2\sqrt{\frac{\ln(2|H_k|/\delta)}{2m}}=L_D(h_i)+\sqrt{2\frac{\ln(2|H_k|/\delta)}{m}}$$

This is true $\forall h_i$, therefore we can substitute $L_D(h_i)$ with $\min_{\mathrm{h_i}\in H_k}L_D(h_i)$. Consequently, we have that with probability at least $1-\delta$, $L_D(h^*)\leq\min_{\mathrm{h_i}\in H_k}L_D(h_i)+\sqrt{\frac{2\ln(2|H_k|/\delta)}{m}}$

B. Bounding the generalization error using *model selection*

Using section (A) with $\delta/2$, the validation set of size αm and $H=\{h_1,\dots,h_k\}$ is of size k, therefore we have

$$P\left(L_D(h^*)\leq\min_{h\in H}L_D(h)+\sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}\right)\geq 1-\frac{\delta}{2}$$

Denote the best h in $\min_{h\in H}L_D(h)$ as h_j, so with probability$\geq 1-\delta/2$ we have

$$L_D(h^*)\leq L_D(h_j)+\sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}$$

the training set, on the other hand, is of size $(1-\alpha)m$, and for every $i\in\{1,2,\dots,k\}$ we have

$$P\left(L_D(h_i) \le \min_{\mathrm{h}\in \mathrm{H}_i} L_D(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\frac{4|H_i|}{\delta}}\right) \ge 1-\frac{\delta}{2}$$

If so, with probability$\ge 1-\delta/2$ we have

$$L_D(h_j) \le \min_{\mathrm{h}\in H_j} L_D(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\frac{4|H_j|}{\delta}}$$

in total the probability of the two independent events is at least $(1-\frac{\delta}{2})^2 = 1-\delta+\frac{\delta^2}{4}$, and specifically is at least $1-\delta$

$$L_D(h^*) \le L_D(h_j) + \sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}} \le \min_{\mathrm{h}\in H_j} L_D(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\frac{4|H_j|}{\delta}} + \sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}$$
$$= \min_{\mathrm{h}\in H_k} L_D(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\frac{4|H_j|}{\delta}} + \sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}$$

C. The first case would be when the model selection is better than the standard method – this can be achieved if the best index j is smaller than k, that is $j \ll k$. for such j if we take the size of H_i to be $|H_i| = 2^i$ for every $i \ne j$ and $|H_j| = c$ for some constant c

For the standard model - $L(h^*) \le \min_{h\in H_k} L(h) + \sqrt{\frac{2}{m}\ln(\frac{2^{k+1}}{\delta})}$

For the model selection - $L(h^*) \le \min_{h\in H_k} L(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\frac{c}{\delta}} + \sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}$

From here it is understood that the standard model error is increasing with magnitude $\sim\sqrt{\ln(exponent(k))}$, that is, linear in k though the model selection increases with $\sim\sqrt{\ln(k)}$, so for large k the error for the model selection would be smaller.

If $j = k$ than for the standard method we have $L(h^*) \le \min_{h\in H_k} L(h) + \sqrt{\frac{2}{m}\ln\left(\frac{2|H_k|}{\delta}\right)}$ and for the model selection we get $L(h^*) \le \min_{h\in H_k} L(h) + \sqrt{\frac{2}{(1-\alpha)m}\ln\frac{4|H_k|}{\delta}} + \sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}$. From here we understand that the standard method is preferred because

$$L(h^*) \overset{model\ selection}{\le} \sqrt{\frac{2}{m(1-\alpha)}\ln\frac{4|H_k|}{\delta}} + \underbrace{\sqrt{\frac{2}{\alpha m}\ln\frac{4k}{\delta}}}_{\ge 0}$$

Since the second element is nonnegative, we can consider the first element and see that

$$\frac{1}{m}\frac{2}{1-\alpha}\ln\frac{4|H_k|}{\delta} = \frac{1}{m}\frac{2}{1-\alpha}\left(\ln(4) + \ln\frac{|\mathrm{H_k}|}{\delta}\right) = \frac{1}{m}\frac{2\ln(4)}{1-\alpha} + \frac{1}{m}\frac{2}{1-\alpha}\ln\frac{|\mathrm{H_k}|}{\delta}$$

Now since $\alpha < 1$ the first element is greater than 1, and the second element is greater than $\frac{2}{m}\ln\frac{|H_k|}{\delta}$. In other words, the model selection's error is greater than the standard model.

Orthogonal Design

2.

A. Prove $\hat{w}_\lambda^{ridge} = \frac{\hat{w}^{LS}}{1+\lambda}$:

We have seen that $\hat{w}_\lambda^{ridge} = (X^TX+\lambda I_d)^{-1}X^Ty$. Using the fact that $X^TX = I_d$ we have

$$\hat{w}_\lambda^{ridge} = (I_d+\lambda I_d)^{-1}X^Ty = \left((1+\lambda)I_p\right)^{-1}X^Ty = \frac{I_p}{1+\lambda}X^Ty = \frac{1}{1+\lambda}X^Ty = \frac{\hat{w}^{LS}}{1+\lambda}$$

B. Prove $\hat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{har}\ (\hat{w}^{LS}) = \mathbf{1}\left[|\hat{w}^{LS}| - \sqrt{\lambda}\right] \cdot \hat{w}^{LS}$:

We know that the error, given some weights vector $w \in \mathbb{R}^d$, is $||y - Xw||^2$. Since we want to minimize the loss, we can multiply by X^T without worrying that the result will change:

$$\left|\left|y - Xw\right|\right|^2 \overset{when\ minimized}{=} \left|\left|X^T y - X^T Xw\right|\right|^2 \overset{\substack{X^T y = \widehat{w}^{LS} \\ X^T X = I_d}}{=} \left|\left|\widehat{w}^{LS} - w\right|\right|^2 = \sum_{i=1}^{n} (\widehat{w}_i^{LS} - w_i)^2$$

Substituting the above to the subset selection problem $\underset{w\in\mathbb{R}^d}{argmin}(\left|\left|y - Xw\right|\right|^2 + \lambda\left|\left|w\right|\right|_0)$ we have $\underset{w\in\mathbb{R}^d}{argmin}(\sum_{i=1}^{n}[(\widehat{w}_i^{LS} - w_i)^2 + \lambda\left|\left|w_i\right|\right|_0]) = \underset{w\in\mathbb{R}^d}{argmin}(\sum_{i=1}^{n}[(\widehat{w}_i^{LS} - w_i)^2 + \lambda \cdot \mathbf{1}[w_i = 0]])$

From here we understand that if $\lambda \geq (\widehat{w}_i^{LS})^2$ for some $i \in [n]$, then $\sqrt{\lambda} \geq |\widehat{w}_i^{LS}|$, which in turn means that $w_i = 0$. This in correlation to the definition of η. If on the other hand $\lambda < (\widehat{w}_i^{LS})^2$, we get $w_i = \widehat{w}_i^{LS}$. In other words, we have $\widehat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{hard}(\widehat{w}^{LS})$

Regularization

3.

A. starting with $A_\lambda \widehat{w} = A_\lambda \widehat{w}(\lambda = 0)$, We know that $\widehat{w}(\lambda = 0)$ is given by $\underset{w}{argmin}(\left|\left|y - Xw\right|\right|_2^2) = (X^T X)^{-1} X^T y$ (this is the solution with no regularization) therefore $A_\lambda \widehat{w} = (X^T X + \lambda I_d)^{-1}(X^T X) \cdot (X^T X)^{-1} X^T y = (X^T X + \lambda I_d)^{-1} X^T y$. We have seen in class that this form is exactly $\widehat{w}(\lambda)$, so we are done.

B. We will show that $\lambda > 0 \Rightarrow \mathbb{E}[\widehat{w}(\lambda)] \neq w$

$$E[\widehat{w}(\lambda)] \overset{(3.A)}{=} E[A_\lambda \widehat{w}] = E[(X^T X + \lambda I_d)^{-1}(X^T X) \cdot \widehat{w}]$$

Since we are given with a constant X, the expectancy of A_λ is simply A_λ

$$= (X^T X + \lambda I_d)^{-1}(X^T X) E[\widehat{w}] \overset{E[\widehat{w}] = w}{=} (X^T X + \lambda I_d)^{-1}(X^T X) \cdot w$$

Now, if λ would have been 0, then A_λ would be I_d. But since $\lambda > 0$ we have $(X^T X + \lambda I_d)^{-1}(X^T X) \neq I_d$, therefore $(X^T X + \lambda I_d)^{-1}(X^T X) w \neq w$, as needed.

C. Show that $Var\big(\widehat{w}(\lambda)\big) = \sigma^2 A_\lambda (X^T X)^{-1} A_\lambda^T$:

Using the hint, we have $Var(A_\lambda \widehat{w}) = A_\lambda Var(\widehat{w}) A_\lambda^T = A_\lambda \sigma^2 (X^T X)^{-1} A_\lambda^T$

D. Similar to what we have seen in class, let us denote y^* as the true hypothesis, $\bar{y}$ as the expectancy $E[\hat{y}]$ and $\hat{y}$ as the estimation. Under those definitions we can write $E[\left|\left|\hat{y} - y^*\right|\right|^2] = Var[\hat{y}] + bias^2[\hat{y}]$. This is because $Var[\hat{y}] = E[\left|\left|\hat{y} - \bar{y}\right|\right|^2]$ and $bias[\hat{y}] = \left|\left|\bar{y} - y^*\right|\right|^2$.

The true hypothesis y^* is the expectancy of $\widehat{w}$ (with no regularization terms), that is $y^* = E[\widehat{w}]$. Furthermore, our estimator (with regularization) is $\hat{y} = \widehat{w}(\lambda)$ therefore also $\bar{y} = E[\widehat{w}(\lambda)]$. We can now calculate the needed values:

For the variance, we have

$$Var(\lambda) = Tr(Var\big(\widehat{w}(\lambda)\big) = \sigma^2 Tr(A_\lambda (X^T X)^{-1} A_\lambda^T)$$

For simplicity let us denote $(X^T X + \lambda I_d) = X'$, so $A_\lambda = X'^{-1}(X^T X)$ and

$$Var(\lambda) = \sigma^2 Tr\left(X'^{-1}(X^T X)(X^T X)^{-1}\big(X'^{-1}(X^T X)\big)^T\right) = \sigma^2 Tr(X'^{-1}(X^T X) X'^{-1})$$

Now, deriving the variance in terms of X' (Trace is a linear operation; hence it commutes with the derivative)

$$\frac{dVar(\lambda)}{dX'}|_{\lambda=0} = \sigma^2 Tr\left(\frac{d}{dX'}(X'^{-1}(X^T X) X'^{-1})\right)_{\lambda=0} = -2\sigma^2 (X^T X)^{-1}(X^T X)^{-1}$$

Now deriving in terms of λ:

$$\frac{dVar(\lambda)}{d\lambda}|_{\lambda=0} = Tr\left(\frac{dVar(\lambda)}{dX'}\frac{dX'}{d\lambda}\right) = -2\sigma^2 Tr((X^TX)^{-1}(X^TX)^{-1})$$

Which is a negative value.

And for the bias

$$bias(\lambda) = ||\bar{y} - y^*|| = ||E[\widehat{w}(\lambda)] - E[\widehat{w}]|| = ||E[A_\lambda w] - E[\widehat{w}]|| = ||A_\lambda E[w] - w||$$
$$= ||A_\lambda w - w|| = ||(A_\lambda - I)w||$$

If so,

$$bias^2(\lambda) = ||(A_\lambda - I)w||^2 = w^T(A_\lambda - I)^T(A_\lambda - I)w$$

Or in terms of X'

$$bias^2(\lambda) = w^T(X'^{-1}(X^TX) \ - I)^T(X'^{-1}(X^TX) \ - I)w$$

Again, deriving (and using summation syntax)

$$\frac{dbias^2}{d\lambda}|_{\lambda=0} = \frac{d}{d\lambda}\left(\sum_i\left(\sum_j X'_{ij}w_j\right)^2\right)|_{\lambda=0}$$
$$= 2\sum_i\left(\sum_j X'_{ij}w_j\right)_{\lambda=0} \cdot \frac{d}{d\lambda}\left(\sum_j X_{ij}w_j\right)_{\lambda=0} = 0$$

So, the MSE is given by

$$MSE = bias^2 + var = w^T(A_\lambda - I)^T(A_\lambda - I)w + \sigma^2 Tr(A_\lambda(X^TX)^{-1}A_\lambda^T$$

And the derivative is the sum of derivatives, which is < 0, simply because the derivative of the bias squared is zero, and the derivative of the variance is negative.

E. We know that the linear model with no regularization is the case where $\lambda = 0$. Now, since we have found in (3.D) that for some $\lambda > 0$, $MSE(\lambda) < 0$, we conclude that such λ satisfies $MSE(\lambda) < MSE(0)$. In other words, using regularization we have decreased the error – which is awesome.

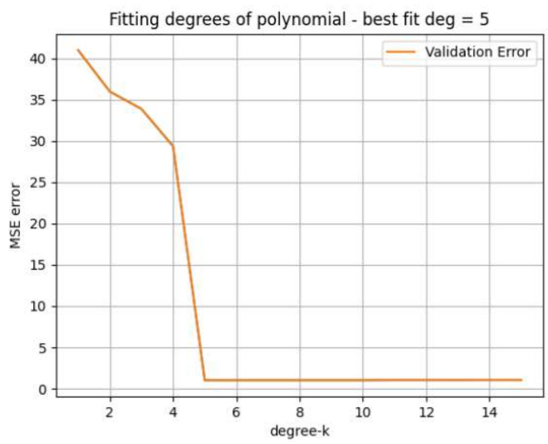
<u>Practical Questions</u>

k-Fold Cross Validation on Polynomial Fitting

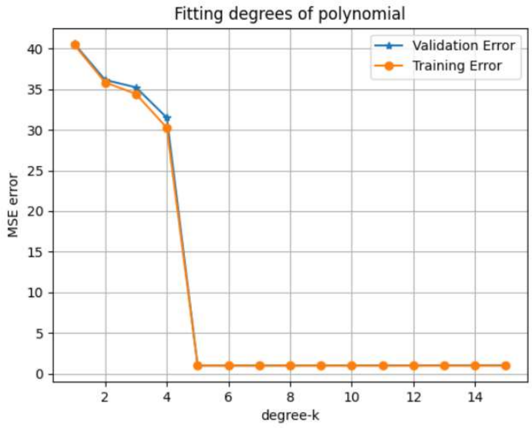
4.

A,B,C & D: code

E. The first graph is for 2-Fold (where each data point was only used once, either for training or for validation):



The second graph is a proper 5-Fold:



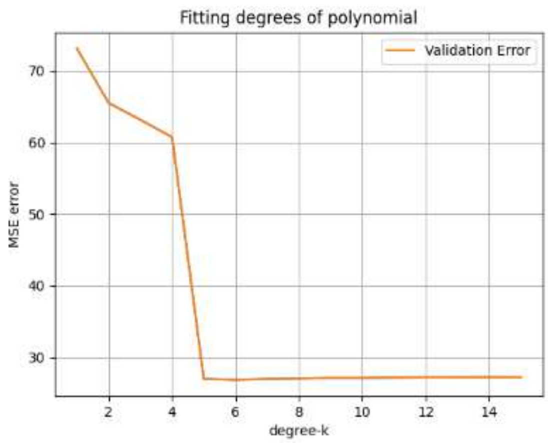
We can see that in both cases $d^* = 5$, which is what we initially expected when fitting the polynomial $f(x) = (x+3)(x+2)(x+1)(x-1)(x-2)$

F. Code

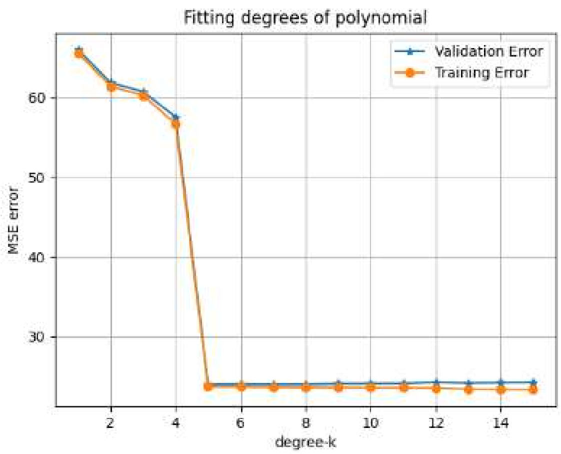
G. When calculating the test error, we see that it is similar to the errors we have encountered in previous items. From here can conclude that k-Fold cross validation is a useful tool when it comes to fitting polynomials.

H. Repeating the process for $\varepsilon \sim \mathcal{N}(0{,}5)$ rather than $\mathcal{N}(0{,}1)$:

2-Fold:



5-Fold:



We can see that for $\sigma = 5$, the data is more prone to overfitting. This can be seen for degrees > 5, in which the 5-Fold validation error increases slightly. Having said that, the final result of $d^* = 5$ still remains, thus we can conclude that even for a noisier data, k-Fold-CV is still a reliable option.

k-Fold and Regularization

5.

A. Code

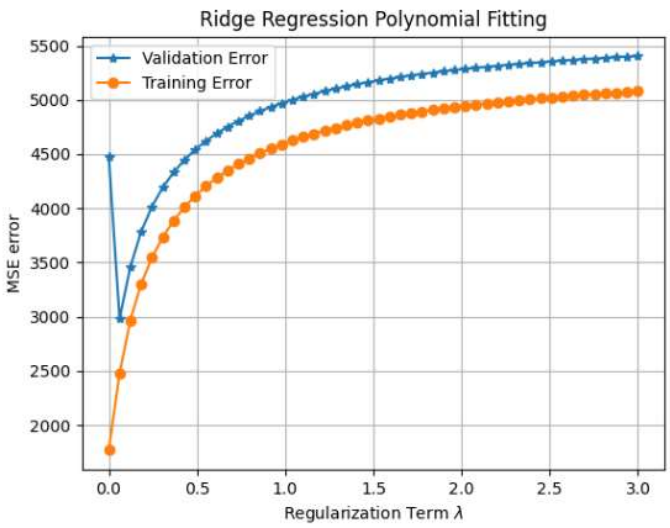
B. Code

C. i. Code

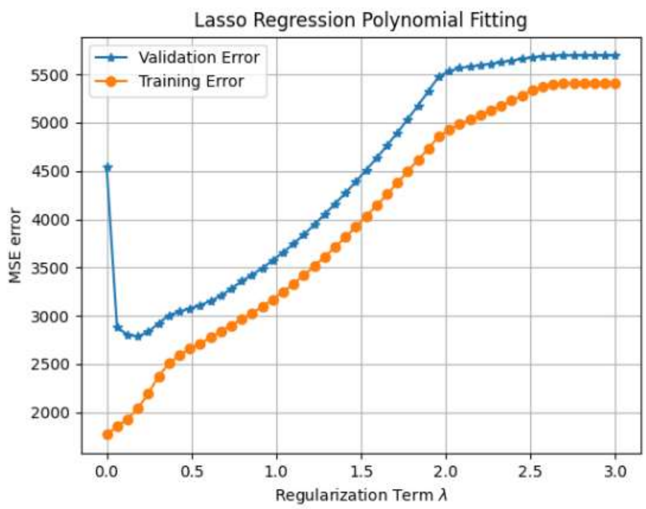
ii. I have chosen the range to be $\lambda \in [0.00001{,}3]$ because on one hand, we can see the behavior with small regularization (λ~0) and expect the model to act similar to what we have previously discussed. The high values of λ (where $\lambda > 1$) are to see some heavy regularization terms effect the error.

D.

plotting the MSE error for Ridge:



Plotting the MSE error for Lasso:



E. The best λ regularization term for both regressions is the one that minimizes the MSE error, which in our case was

Lasso: 0.18368285714285715

Ridge: 0.06123428571428571

We can see that the regularization term that provided the minimum MSE is rather small, which indicates that the contribution of such term may not be of utter importance.

F and G. the error on the Test-Set turns out to be

Ridge: 3211.228315328465

Lasso: 3393.866002033245

Linear: 3612.2496883248987

As mentioned in section (E), the regularization term λ was rather small. Having said that, we can still see some difference when calculating the error – it is clear that the linear regression model (with no regularization) provided a little higher error over the regularized models, and Ridge performed slightly better than lasso