Introduction to Machine Learning (67577)

Exercise 5 - Validation, Feature Selection and Regularization

Dor Roter
208772251

June 19, 2021

1 Validation

1. (a) Let us first note that the given loss function is bounded by 1. let us denote it by l, then for all $h \in \mathcal{H}_k$:

$$X_i = l\left(h\left(x_i\right), y_i\right) \in [0, 1]$$

and:

$$L_{S_{all}}\left(h\right) = \frac{1}{m}\sum_{i=1}^{m} l\left(h\left(x_i\right), y_i\right) = \frac{1}{m}\sum_{i=1}^{m} X_i$$
$$L\left(h\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[l\left(h\left(x\right), y\right)\right] \underset{X\text{'s iid}}{=} \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right]$$

Where X_i is a bounded i.i.d random variable (as h is selection is dependant on $S \sim \mathcal{D}^m$ and so is the selection of $(x, y) \sim \mathcal{D}$).
Let $\delta \in (0, 1)$ using hoeffding inequality let us find an ϵ for which S is ϵ-represnetive of $\mathcal{H}_k$.
Firstly for all $h \in \mathcal{H}_k$ it holds that the probability that S_{all} is not ϵ-representive is bounded such that:

$$\mathbb{P}\left[\left|L_{S_{all}}\left(h\right) - L\left(h\right)\right| \geq \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

Therfore the probablity for **any** $h \in \mathcal{H}_k$ to be not ϵ-representive is bounded by the union bound (as $\mathcal{H}_k$ is finite):

$$\mathbb{P}\left[\exists h \in \mathcal{H}_k \,|\left|L_{S_{all}}\left(h\right) - L\left(h\right)\right| \geq \epsilon\right] \underset{\text{union bound}}{\leq} \left|\mathcal{H}_k\right| \cdot \underset{h' \in \mathcal{H}_k}{max}\left(\mathbb{P}\left[\left|L_{S_{all}}\left(h'\right) - L\left(h'\right)\right| \geq \epsilon\right]\right) \leq 2\left|\mathcal{H}_k\right| e^{-2m\epsilon^2}$$

Finally we can find the ϵ that bounds $\mathcal{D}^m\left[S\,||L_{S_{all}}(h)-L(h)|\geq\epsilon\right]\leq 2|\mathcal{H}_k|\,e^{-2m\epsilon^2}\leq\delta$:

$$\begin{aligned}
&2|\mathcal{H}_k|\,e^{-2m\epsilon^2}\leq\delta\\
&\Leftrightarrow ln\left(2|\mathcal{H}_k|\right)+ln\left(e^{-2m\epsilon^2}\right)\leq ln(\delta)\\
&\Leftrightarrow ln\left(2|\mathcal{H}_k|\right)-ln(\delta)\leq 2m\epsilon^2\\
&\Leftrightarrow \frac{ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}\leq\epsilon^2\\
&\underset{\frac{1}{\delta}>1}{\Leftrightarrow}\epsilon\geq\sqrt{\frac{ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}}
\end{aligned}$$

Therefore for $\epsilon=\sqrt{\frac{ln(2|\mathcal{H}_k|/\delta)}{2m}}$ it holds for any δ that:

$$\begin{aligned}
&\mathbb{P}\left[|L_{S_{all}}(h)-L(h)|>\sqrt{\frac{ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}}\right]\leq\delta\\
\Leftrightarrow&\mathbb{P}\left[|L_{S_{all}}(h)-L(h)|\leq\sqrt{\frac{ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}}\right]\geq 1-\delta
\end{aligned}$$

So it holds that with probability of at least $(1-\delta)$ S_{all} is $\sqrt{\frac{ln(2|\mathcal{H}_k|/\delta)}{2m}}$-representive of $\mathcal{H}_k$ and so as we have shown in recitation, for $h^*\in ERM_{\mathcal{H}_k}$ it holds that:

$$L(h^*)\leq\underset{h\in\mathcal{H}_k}{min}\,L(h)+2\sqrt{\frac{ln\left(2|\mathcal{H}_k|/\delta\right)}{2m}}=\underset{h\in\mathcal{H}_k}{min}\,L(h)+\sqrt{\frac{2\cdot ln\left(2|\mathcal{H}_k|/\delta\right)}{m}}$$

□

(b) Let us use the previos inequality for both the second and third step of model selection.
Let $\delta,\alpha\in(0,1)$.
For the second step, let $i\in[k]$ then $h_i\in ERM_{\mathcal{H}_i}(S)$ where S size is $(1-\alpha)m$, so:

$$L(h_i)\leq\underset{h\in\mathcal{H}_i}{min}\,L(h)+\sqrt{\frac{2\cdot ln\left(2|\mathcal{H}_i|/(\delta/2)\right)}{(1-\alpha)m}}=\underset{h\in\mathcal{H}_i}{min}\,L(h)+\sqrt{\frac{2}{(1-\alpha)m}ln\left(\frac{4|\mathcal{H}_i|}{\delta}\right)}$$

Similarly for the third step with $|V|=\alpha m$ and $|\mathcal{H}|=k$:

$$L(h^*)\leq\underset{h\in\mathcal{H}}{min}L(h)+\sqrt{\frac{2\cdot ln\left(4|\mathcal{H}|/\delta\right)}{\alpha m}}=\underset{h\in\mathcal{H}}{min}L(h)+\sqrt{\frac{2}{\alpha m}ln\left(\frac{4k}{\delta}\right)}$$

Both with probability of at least $1-\frac{\delta}{2}$.
As noted in the question, if $\overline{h}=argmin_{h\in\mathcal{H}_k}L(h)\in\mathcal{H}_j$ then for $h_j\in\mathcal{H}$:

$$\underset{h\in\mathcal{H}}{min}L(h)\underset{min}{\leq}L(h_j)\underset{above}{\leq}\underset{h\in\mathcal{H}_j}{min}\,L(h)+\sqrt{\frac{2}{(1-\alpha)m}ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}$$

Finally, as the two events are independent, the probability for both is $\left(1-\frac{\delta}{2}\right)^2=1-\delta+\delta^2>1-\delta$, so clearly with probability of at least $1-\delta$ it holds that:

$$L(h^*) \le \underset{h\in\mathcal{H}}{min} L(h) + \sqrt{\frac{2}{\alpha m} ln\left(\frac{4k}{\delta}\right)} \le \underset{h\in\mathcal{H}_j}{min} L(h) + \sqrt{\frac{2}{\alpha m} ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m} ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)}$$

Since $argmin_{h\in\mathcal{H}_k} L(h) \in \mathcal{H}_j \subseteq \mathcal{H}_{j+1} \subseteq \cdots \subseteq \mathcal{H}_k$ it holds that $\underset{h\in\mathcal{H}_j}{min} L(h) = \underset{h\in\mathcal{H}_k}{min} L(h)$, thus finally:

$$L(h^*) \le \underset{h\in\mathcal{H}_j}{min} L(h) + \sqrt{\frac{2}{\alpha m} ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m} ln\left(\frac{4|\mathcal{H}_j|}{\delta}\right)} = \underset{h\in\mathcal{H}_k}{min} L(h) + \sqrt{\frac{2}{\alpha m} ln\left(\frac{4k}{\delta}\right)} + \sqrt{\frac{2}{(1-\alpha)m}} ln(4|$$

$\square$

(c) If $\mathcal{H}_j = \mathcal{H}_k$ then clearly

$$\sqrt{\frac{2ln(2|\mathcal{H}_k|/\delta)}{m}} < \sqrt{\frac{2ln(4|\mathcal{H}_k|/\delta)}{(1-\alpha)m}} = \sqrt{\frac{2ln(4|\mathcal{H}_j|/\delta)}{(1-\alpha)m}}$$

And then the standard method is bounded tighter than the model selection method:

$$L(h^*) \le \underset{h\in\mathcal{H}_k}{min} L(h) + \sqrt{\frac{2ln(2|\mathcal{H}_k|/\delta)}{m}} \le \underset{h\in\mathcal{H}_k}{min} L(h) + \sqrt{\frac{2ln(4|\mathcal{H}_j|/\delta)}{(1-\alpha)m}} \le \underset{h\in\mathcal{H}_k}{min} L(h) + \sqrt{\frac{2ln(4k/\delta)}{\alpha m}} + \sqrt{\frac{2ln(4|\mathcal{H}_j|/}{(1-\alpha)m}}$$

Next, let $|\mathcal{H}_i| = 2^i$:

$$\frac{\sqrt{\frac{2ln(4k/\delta)}{\alpha m}} + \sqrt{\frac{2ln(4|\mathcal{H}_j|/\delta)}{(1-\alpha)m}}}{\sqrt{\frac{2ln(2|\mathcal{H}_k|/\delta)}{m}}} = \left(\sqrt{\frac{2ln(4k/\delta)}{\alpha m}} + \sqrt{\frac{2ln(4|\mathcal{H}_j|/\delta)}{(1-\alpha)m}}\right) \cdot \frac{\sqrt{m}}{\sqrt{2ln(2|\mathcal{H}_k|/\delta)}} =$$

$$\sqrt{\frac{m\cdot 2ln(4k/\delta)}{2ln(2|\mathcal{H}_k|/\delta)\cdot \alpha m}} + \sqrt{\frac{m\cdot 2ln(4|\mathcal{H}_j|/\delta)}{2ln(2|\mathcal{H}_k|/\delta)\cdot(1-\alpha)m}} = \sqrt{\frac{ln(4k/\delta)}{ln(2|\mathcal{H}_k|/\delta)\cdot\alpha}} + \sqrt{\frac{ln(4|\mathcal{H}_j|/\delta)}{ln(2|\mathcal{H}_k|/\delta)\cdot(1-\alpha)}} =$$

$$\sqrt{\frac{ln(4k/\delta)}{(ln(2^{k+1})-ln(\delta))\cdot\alpha}} + \sqrt{\frac{ln(2^{j+2}/\delta)}{(ln(2^{k+1})-ln(\delta))\cdot(1-\alpha)}} \underset{\delta<1}{\le} \sqrt{\frac{ln(4k)-ln(\delta)}{ln(2^{k+1})\cdot\alpha}} + \sqrt{\frac{ln(2^{j+2})-ln(\delta)}{ln(2^{k+1})\cdot(1-\alpha)}} =$$

$$= \sqrt{\frac{O(ln(k))}{O(k)}} + \sqrt{\frac{O(j)}{O(k)}}$$

And thus especially when j is constant and $k\to\infty$ (any case where j is sufficently smaller than k):

$$\frac{\sqrt{\frac{2ln(4k/\delta)}{\alpha m}} + \sqrt{\frac{2ln(4|\mathcal{H}_j|/\delta)}{(1-\alpha)m}}}{\sqrt{\frac{2ln(2|\mathcal{H}_k|/\delta)}{m}}} < 1 \Leftrightarrow \sqrt{\frac{2ln(4k/\delta)}{\alpha m}} + \sqrt{\frac{2ln(4|\mathcal{H}_j|/\delta)}{(1-\alpha)m}} < \sqrt{\frac{2ln(2|\mathcal{H}_k|/\delta)}{m}}$$

And so the model selection method offers better bounds.

2. (a) We know $\hat{w}_\lambda^{ridge} = \left(X^TX+\lambda I\right)^{-1}X^Ty$ is the closed solution for the ridge optimization, and $\hat{w}^{LS} = \left(X^TX\right)^\dagger X^Ty = X^Ty$ is the closed solution the regular regression problem, and so:

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

(b) Firstly $\hat{w}^{LS} = X^T y$ under orthogonal design.

$$\hat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{hard}$$

$$\begin{aligned}\hat{w}_\lambda^{subset} = argmin_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \left|\left|w_0 1 + Xw - y\right|\right|^2 + \lambda \left|\left|w\right|\right|_0 = \\ argmin_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \left|\left|X^T \left(w_0 1 + Xw - y\right)\right|\right|^2 + \lambda \left|\left|w\right|\right|_0 = \\ argmin_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \left|\left|X^T w_0 1 + X^T X w - X^T y\right|\right|^2 + \lambda \left|\left|w\right|\right|_0 = \\ argmin_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \left|\left|X^T w_0 1\right|\right|^2 + \left|\left|w - \hat{w}^{LS}\right|\right|^2 + \lambda \left|\left|w\right|\right|_0 =\end{aligned}$$

Now for each $i \in [d]$ it holds that:

$$argmin_{w_i \in \mathbb{R}} \left(\left(w_i - \hat{w}_i^{LS}\right)^2 + \lambda \left|\left|w_i\right|\right|_0\right) = \begin{cases} argmin_{w_i \in \mathbb{R}} \left(\left(w_i - \hat{w}_i^{LS}\right)^2 + \lambda\right) = \hat{w}_i^{LS} & \left(\hat{w}_i^{LS}\right)^2 > \lambda \Leftrightarrow \left|\hat{w}_i^{LS}\right| > \sqrt{\lambda} \\ argmin_{w_i \in \mathbb{R}} \left(\left(\hat{w}_i^{LS}\right)^2\right) = 0 & \left(\hat{w}_i^{LS}\right)^2 \leq \lambda \Leftrightarrow \left|\hat{w}_i^{LS}\right| \leq \sqrt{\lambda} \end{cases}$$

As w_0 is minimized separatly, $argmin_{w_0 \in \mathbb{R}, w \in \mathbb{R}^d} \left|\left|w_0 1 + Xw - y\right|\right|^2 + \lambda \left|\left|w\right|\right|_0$ can be computed by finding for each index its minimizer, and thus in-fact $\hat{w}_\lambda^{subset} = \eta_{\sqrt{\lambda}}^{hard}$. $\square$

3. (a) First as $X^T X$ is invertible, the closed solution to the linear regression is giveb by $\hat{w}(\lambda = 0) = \hat{w} = \left(X^T X\right)^\dagger X^T y = \left(X^T X\right)^{-1} X^T y$, while the solution to the ridge regression is given by $\hat{w}(\lambda) = argmin_w \left(\left|\left|y - Xw\right|\right|_2^2 + \right.$ $\left(X^T X + \lambda I\right)^{-1} X^T y$, therefore:

$$\begin{aligned}A_\lambda \hat{w} = \left(X^T X + \lambda I\right)^{-1} \left(X^T X\right) \left(X^T X\right)^{-1} X^T y = \\ = \left(X^T X + \lambda I\right)^{-1} X^T y = \hat{w}(\lambda)\end{aligned}$$

$\square$

(b) Since A_λ is a non-random matrix (defined by X, λ which are provided), applying it to $\hat{w}$ is applying a linear transformation to it, thus by the expected value's linearity it holds that:

$$\mathbb{E}\left[\hat{w}(\lambda)\right] = \mathbb{E}\left[A_\lambda \hat{w}\right] = A_\lambda \mathbb{E}\left[\hat{w}\right] = A_\lambda w = \left(X^T X + \lambda I\right)^{-1} \left(X^T X\right) w$$

Therefore for any $\lambda \neq 0$ it holds that $\mathbb{E}\left[\hat{w}(\lambda)\right] \neq w$. $\square$

(c) Using the hints:

$$Var\left(\hat{w}(\lambda)\right) = Var\left(A_\lambda \hat{w}\right) = A_\lambda Var\left(\hat{w}\right) A_\lambda^T = A_\lambda \sigma^2 \left(X^T X\right)^{-1} A_\lambda^T = \sigma^2 A_\lambda \left(X^T X\right)^{-1} A_\lambda^T$$

(d) As we have seen previously, the MSE can be broken up into a bias-variance decomposition:

$$MSE\left(w, \hat{w}\right) = \mathbb{E}\left[\left|\left|\hat{w}(\lambda) - w\right|\right|^2\right] = Var\left(\hat{w}(\lambda)\right) + bias^2\left(\hat{w}(\lambda)\right)$$

We have shown that $\mathbb{E}[\hat{w}] = w$, and so

$$bias(\lambda) = \mathbb{E}[\hat{w}(\lambda) - w] = \mathbb{E}[\hat{w}(\lambda)] - w = (A_\lambda - I)w \Rightarrow bias^2(\lambda) = ||(A_\lambda - I)w||^2$$
$$Var(\lambda) = Var(\hat{w}(\lambda)) = \sigma^2 A_\lambda \left(X^T X\right)^{-1} A_\lambda^T$$

Since $(A_\lambda - I)|_{\lambda=0} = 0$, using the chain rule:

$$\frac{\partial}{\partial\lambda} bias^2(\lambda)|_{\lambda=0} = ||(A_\lambda - I)w||^2 = 2\,||(A_\lambda - I)w||\,|_{\lambda=0}\left(\frac{\partial}{\partial\lambda}(A_\lambda - I)w\right)|_{\lambda=0} = 0 \cdot \left(\frac{\partial}{\partial\lambda}(A_\lambda - I)w\right)|_{\lambda=0} = 0$$

Also since $A_\lambda|_{\lambda=0} = \left(X^T X + \lambda I\right)^{-1} X^T X|_{\lambda=0} = \left(X^T X\right)^{-1} X^T X = I$, Using the chain rule we get:

$$\frac{\partial}{\partial\lambda} A_\lambda|_{\lambda=0} = \left(X^T X\right)\frac{\partial}{\partial\lambda}\left(X^T X + \lambda I\right)^{-1}|_{\lambda=0} = -\left(X^T X\right)\left(X^T X + \lambda I\right)^{-2}\frac{\partial}{\partial\lambda}\left(X^T X + \lambda I\right)|_{\lambda=0} =$$
$$= -\left(X^T X\right)\left(X^T X\right)^{-2} = -\left(X^T X\right)^{-1}$$

$$\frac{\partial}{\partial\lambda} Var(\lambda)|_{\lambda=0} = \frac{\partial}{\partial\lambda}\sigma^2 A_\lambda \left(X^T X\right)^{-1} A_\lambda^T|_{\lambda=0} = 2\sigma^2\left(X^T X\right)^{-1} A_\lambda^T \cdot \left(\frac{\partial}{\partial\lambda} A_\lambda\right)|_{\lambda=0} =$$
$$= 2\sigma^2\left(X^T X\right)^{-1} \cdot -\left(X^T X\right)^{-1} = -2\sigma^2\left(X^T X\right)^{-2}$$

As $X^T X$ is invertible and symetric, it is a PSD, and so $-2\sigma^2\left(X^T X\right)^{-2} < 0$, therefore:

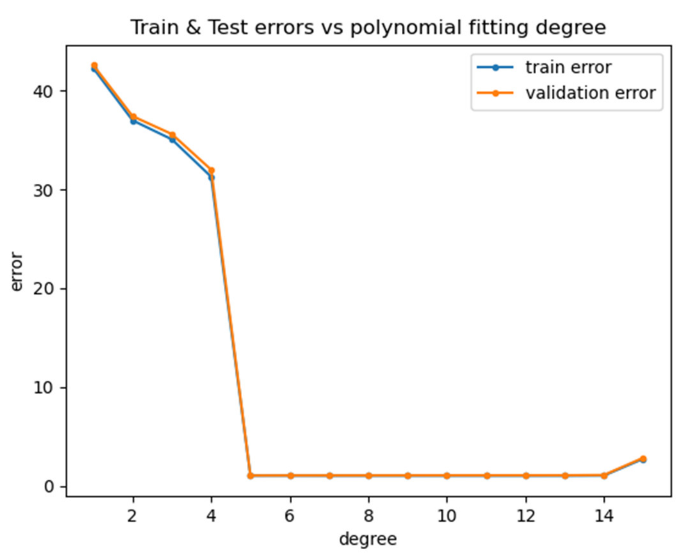
$$\frac{\partial}{\partial\lambda} MSE(\lambda)|_{\lambda=0} = \frac{\partial}{\partial\lambda} Var(\lambda)|_{\lambda=0} + \frac{\partial}{\partial\lambda} bias^2(\lambda)|_{\lambda=0} = -2\sigma^2\left(X^T X\right)^{-2} + 0 < 0$$

□

(e) As we have shown in the last question that $\frac{\partial}{\partial\lambda} MSE(0) < 0$, by definition there exists $\lambda > 0$ such that $\frac{MSE(\lambda) - MSE(0)}{\lambda - 0} < 0 \Rightarrow MSE(\lambda) < MSE(0)$. □

4. (e) As f is a polynimial of the 5th degree, we note $d^* = 5$ which is to be expected as there is not a lot of noise applied to the dataset.

Figure 1: 5-fold validation errors



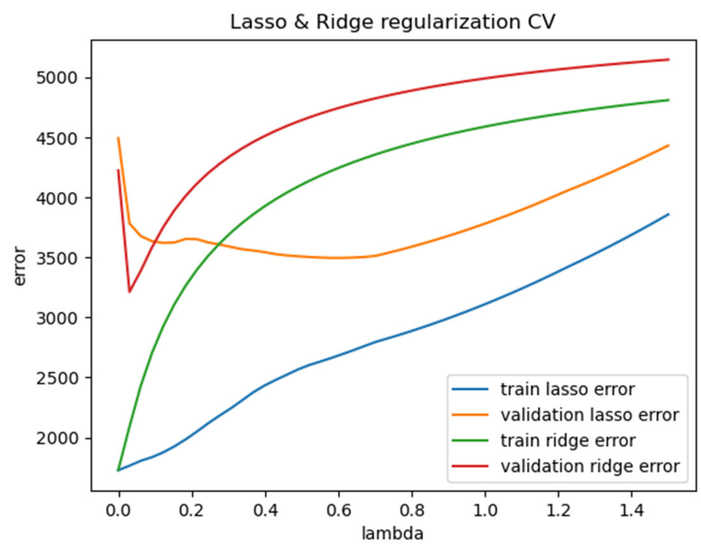
(g) The test error for h^* is around 1.03, which is similar to the cross-validation minimum.

(h) In this case, there is a lot of noise in the dataset, and as a result of this the polynomial fitting model tends to prefer higher degree polynomials as those provide a better traininng error, but thanks to cross-validation we still manage to filter this bias of the model towards overfitting, and find the best fit to be $d^* = 5$.

5. (c) As we are aiming to test for the best regularization parameter (lambda), We would like to see how none-regularized models all the way up to heavliy regularized models fair one against the other. As such, I have elected the range of possible values for lambda to be of linearliy spaced values between zero and 2 (so we have both larger, and smaller than 1 lambda values to compare).

(d)

Figure 2: Training & Validation errors over λ



(g) The best results were achieved by the ridge regressor, it seems that a small amount of regularization was beneficial when comparing the ridge model to the un-regularized linear regressor.