CSE 417T: Homework 3

Hangxiao Zhu

October 18, 2022

Problem 1.

(a) By definition of the weight decay regularizer, we have

$$E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda \vec{w}^T \vec{w}$$

After taking derivative with respect to $\vec{w}$, we have

$$\nabla E_{aug}(\vec{w}) = \nabla E_{in}(\vec{w}) + 2\lambda \vec{w}$$

Therefore, the update rule can be written as

$$\begin{aligned}
\vec{w}(t+1) &= \vec{w}(t) - \eta \nabla E_{aug}(\vec{w}(t)) \\
&= \vec{w}(t) - \eta(\nabla E_{in}(\vec{w}(t)) + 2\lambda \vec{w}(t)) \\
&= (1 - 2\eta\lambda)\vec{w}(t) - \eta \nabla E_{in}(\vec{w}(t))
\end{aligned}$$

(b) By definition of the L_1 regularizer, we have

$$E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \lambda ||\vec{w}||_1$$

Since the gradient of 1-norm is not well-defined at 0, we define a $sign()$ function to address this issue

$$\frac{\partial}{\partial w_i} ||\vec{w}||_1 = sign(w_i) = \begin{cases} +1 & \text{if } w_i > 0 \\ 0 & \text{if } w_i = 0 \\ -1 & \text{if } w_i < 0 \end{cases}$$

After taking derivative of the L_1 regularizer with respect to $\vec{w}$, we have

$$\nabla E_{aug}(\vec{w}) = \nabla E_{in}(\vec{w}) + \lambda \sum_{i=0}^{d} sign(w_i)$$

Therefore, the update rule can be written as

$$\begin{aligned}
\vec{w}(t+1) &= \vec{w}(t) - \eta \nabla E_{aug}(\vec{w}(t)) \\
&= \vec{w}(t) - \eta(\nabla E_{in}(\vec{w}(t)) + \lambda \sum_{i=0}^{d} sign(w_i(t)))
\end{aligned}$$

(c) Report on each of the λ for both L_1 abd L_2 regularizations:

Regularizer	λ	Classification error on test set	Number of 0s in w
L1	0	0.102803738317757	8
L1	0.0001	0.09813084112149532	8
L1	0.001	0.09345794392523364	15
L1	0.005	0.08878504672897196	26
L1	0.01	0.0794392523364486	36
L1	0.05	0.102803738317757	52
L1	0.1	0.13551401869158877	57
L2	0	0.102803738317757	8
L2	0.0001	0.102803738317757	8
L2	0.001	0.09345794392523364	8
L2	0.005	0.09813084112149532	8
L2	0.01	0.09813084112149532	8
L2	0.05	0.11682242990654206	8
L2	0.1	0.12149532710280374	8

Observations based on the results:

- For both regularizations, the classification error decreases and then increases as λ increases. For L_1 regularization, the classification error is smallest when λ is around 0.1, for L_2 regularization, the classification error is smallest when λ is around 0.001.
- For L_1 regularization, the number of zeros in the learned w increases as λ increases, for L_2 regularization, the number of zeros in the learned w keeps the same as λ increases.

Properties of the L_1 regularizer:

- The number of zeros in the learned w increases as λ increases.
- L_1 regularizer helps us discard variables with coefficient zero, so it is useful for feature selection.

Problem 2.

(a) By definition, we have $\vec{w}^T\vec{\Gamma}^T\vec{\Gamma}\vec{w} \leq C$. Since $\sum_{q=0}^{Q} w_q^2 = \vec{w}^T\vec{w} \leq C$, we have $\vec{w}^T\vec{\Gamma}^T\vec{\Gamma}\vec{w} = \vec{w}^T\vec{w}$. Therefore, $\vec{\Gamma}^T\vec{\Gamma} = \vec{I}$. Thus $\vec{\Gamma} = \vec{I}$.

(b)By definition, we have $\vec{w}^T\vec{\Gamma}^T\vec{\Gamma}\vec{w} \leq C$. Since $(\sum_{q=0}^{Q} w_q)^2 = \vec{w}^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \vec{w} \leq C$, we have $\vec{w}^T\vec{\Gamma}^T\vec{\Gamma}\vec{w} = \vec{w}^T \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \vec{w}$. Therefore, $\vec{\Gamma}^T\vec{\Gamma} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$. Thus, $\vec{\Gamma} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$.

Problem 3.

(a) We should not select the learner with minimum validation error. Based on the VC bound equation

$$E_{out}(g_{m^*}^-) \leq E_{val}(g_{m^*}^-) + O(\sqrt{\frac{\ln M}{2K}})$$

we know that the bound is not only depending on $E_{val}(g^-_{m^*})$, but also depending on $O(\sqrt{\frac{\ln M}{2K}})$. For M learners, each learner leads to a unique $O(\sqrt{\frac{\ln M}{2K}})$ value because each learner m has a unique size of their validation set K_m. Therefore, a learner has the smallest $E_{val}(g^-_{m^*})$ does not necessarily mean this learner also has the smallest sum of $E_{val}(g^-_{m^*}) + O(\sqrt{\frac{\ln M}{2K}})$. Thus, the learner with minimum validation error might not generate the tightest VC bound.

(b) Because when all models are validated on the same validation set, each learner will have the same $O(\sqrt{\frac{\ln M}{2K}})$. Therefore, the learner with the smallest $E_{val}(g^-_{m^*})$ is guaranteed to have the smallest sum of $E_{val}(g^-_{m^*}) + O(\sqrt{\frac{\ln M}{2K}})$. Thus, the learner with minimum validation error will generate the tightest VC bound.

(c) According to Hoeffding's Inequality, for each m we have

$$
\begin{aligned}
&\mathbb{P}[|E_{out}(m) - E_{val}(m)| > \epsilon] \leq 2e^{-2\epsilon^2 K_m} \\
\Rightarrow\ &\mathbb{P}[E_{out}(m) - E_{val}(m) > \epsilon] \leq e^{-2\epsilon^2 K_m} \\
\Rightarrow\ &\mathbb{P}[E_{out}(m) > E_{val}(m) + \epsilon] \leq e^{-2\epsilon^2 K_m}
\end{aligned}
$$

Since

$$
\begin{aligned}
\mathbb{P}[E_{out}(m^*) > E_{val}(m^*) + \epsilon] &\leq \mathbb{P}[(E_{out}(m_1) > E_{val}(m_1) + \epsilon) \\
&\quad \text{or } (E_{out}(m_2) > E_{val}(m_2) + \epsilon) \\
&\quad \text{or } \cdots \\
&\quad \text{or } (E_{out}(m_M) > E_{val}(m_M)] \\
&\leq \mathbb{P}[E_{out}(m_1) > E_{val}(m_1) + \epsilon] \\
&\quad + \mathbb{P}[E_{out}(m_2) > E_{val}(m_2) + \epsilon] \\
&\quad + \cdots \\
&\quad + \mathbb{P}[E_{out}(m_M) > E_{val}(m_M) + \epsilon] \\
&\leq \sum_{m=1}^{M} e^{-2\epsilon^2 K_m}
\end{aligned}
$$

Since we have the average validation set size

$$
\kappa(\epsilon) = -\frac{1}{2\epsilon^2} \ln(\frac{1}{M} \sum_{m=1}^{M} e^{-2\epsilon^2 K_m})
$$

We can deduce that

$$
\begin{aligned}
Me^{-2\epsilon^2 \kappa(\epsilon)} &= Me^{\ln(\frac{1}{M}\sum_{m=1}^{M} e^{-2\epsilon^2 K_m})} \\
&= M\frac{\sum_{m=1}^{M} e^{-2\epsilon^2 K_m}}{M} \\
&= \sum_{m=1}^{M} e^{-2\epsilon^2 K_m}
\end{aligned}
$$

Therefore, we have

$$
\mathbb{P}[E_{out}(m^*) > E_{val}(m^*) + \epsilon] \leq Me^{-2\epsilon^2 \kappa(\epsilon)}
$$

Problem 4.

(a)According to Hoeffding's Inequality, we have

$$\mathbb{P}[|E_{out} - E_{in}| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$
$$\Rightarrow \mathbb{P}[|E_{in} - E_{out}| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

i The problem here is data snooping. Since we are using 50 years of data, and the S&P 500 stocks were selected by looking at the whole data set, when we use $M = 500$ to decide whether the stock we picked is profitable, we are actually underestimate the M.

ii According to (i), we should use $M = 50000$ to do the estimation. Using the hoeddfing bound, we have

$$\mathbb{P}[|E_{in} - E_{out}| > 0.02] \leq 2 \times 50000 \times e^{-2\times 12500\times 0.02^2} \approx 4.54$$

(b)

i The problem here is data snooping. Since we are using 50 years of data, and the S&P 500 stocks were selected by looking at the whole data set, we cannot generalize this conclusion to the entire data set.

ii Since our analysis of the performance of buy and hold trading is only based on today's S&P 500 stocks, we can say that in practice, the performance of 'buy and hold' strategy will be worse than our estimation.