Homework #1

CSE 546: Machine Learning
Ray Chen
Collaborator: Kevin Shao

October 20, 2022

B1:

- **Part a:** Proof
For Trian error:

$$\begin{aligned}
\mathbb{E}_{\text{train}}\left[\hat{\epsilon}_{\text{train}}(\hat{f}_{\text{train}})\right] &= \mathbb{E}\big[\frac{1}{N_{train}}\sum_{(x,y)\in S_{trian}}(f(x)-y)^2\big] \\
&= \frac{1}{N_{train}}\sum_{(x,y)\in S_{trian}}\mathbb{E}(f(x)-y)^2 \\
&= N_{train}\frac{1}{N_{train}}\sum_{(x,y)\sim D}\mathbb{E}(f(x)-y)^2 \\
&= \epsilon(f)
\end{aligned}$$

Similarly, for test error

$$\begin{aligned}
\mathbb{E}_{\text{test}}\left[\hat{\epsilon}_{\text{test}}(\hat{f}_{\text{test}})\right] &= \mathbb{E}\big[\frac{1}{N_{test}}\sum_{(x,y)\in S_{test}}(\hat{f}(x)-y)^2\big] \\
&= \frac{1}{N_{test}}\sum_{(x,y)\in S_{test}}\mathbb{E}(\hat{f}(x)-y)^2 \\
&= N_{test}\frac{1}{N_{test}}\sum_{(x,y)\sim D}\mathbb{E}(\hat{f}(x)-y)^2 \\
&= \epsilon(\hat{f})
\end{aligned}$$

- **Part b:** Brief Explanation (3-5 sentences)
No, it will not be true with regard to the training loss. Also, $\mathbb{E}_{\text{train}}\left[\hat{\epsilon}_{\text{train}}(\hat{f}_{\text{train}})\right] \neq \epsilon(\hat{f})$. Because x_i and y_i are not independent to the $\hat{f}$.

- **Part c:** Proof

$$\begin{aligned}
\mathbb{E}_{\text{train}}\left[\hat{\epsilon}_{\text{train}}(\hat{f}_{\text{train}})\right] &= \sum_{f\in\mathcal{F}}\mathbb{E}_{\text{train}}\left[\hat{\epsilon}_{\text{train}}(\hat{f}_{\text{train}})\right]\mathbb{P}_{\text{train}}(\hat{f}_{\text{train}}=f) \\
&= \sum_{f\in\mathcal{F}}\mathbb{E}_{\text{test}}\left[\hat{\epsilon}_{\text{test}}(f)\right]\mathbb{P}_{\text{train}}(\hat{f}_{\text{train}}=f) \\
&= \mathbb{E}_{\text{train,test}}\left[\hat{\epsilon}_{\text{test}}(\hat{f}_{\text{train}})\right]
\end{aligned}$$

B2:

- **Part a:** 1-2 sentences Intuitively, for a small m, it should have low bias and high variance; for a large m, it should have high bias and low variance. According to formula of bias-variance decomposition at x_i:

$$\mathbb{E}\left[(\widehat{f}_m(x_i) - f(x_i))^2\right] = \underbrace{(\mathbb{E}[\widehat{f}_m(x_i)] - f(x_i))^2}_{\text{Bias}^2(x_i)} + \underbrace{\mathbb{E}\left[(\widehat{f}_m(x_i) - \mathbb{E}[\widehat{f}_m(x_i)])^2\right]}_{\text{Variance}(x_i)}$$

- **Part b:** Proof

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\big(\mathbb{E}[\hat{f}_m(x_i)] - f(x_i)\big)^2 =& \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\sum_{j=1}^{n/m} c_j\mathbf{1}\big] - f(x_i)\Big)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}\Big(\mathbb{E}\big[\sum_{j=1}^{n/m}\big(\frac{1}{m}\sum_{k=(j-1)m+1}^{jm} y_k\big)\mathbf{1}\big] - f(x_i)\Big)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}\big(\sum_{j=1}^{n/m}\big(\frac{1}{m}\sum_{k=(j-1)m+1}^{jm}\mathbb{E}\big[f(x_k) + \epsilon_k\big]\big)\mathbf{1} - f(x_i)\big)^2 \\
=& \frac{1}{n}\sum_{i=1}^{n}\big(\sum_{j=1}^{n/m}\bar{f}^{(j)}\mathbf{1} - f(x_i)\big)^2 \\
=& \frac{1}{n}\sum_{j=1}^{n/m}\sum_{i=(j-1)m+1}^{jm}\big(\bar{f}^{(j)} - f(x_i)\big)^2
\end{aligned}$$

- **Part c:** Proof

$$
\begin{aligned}
\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{f}_m(x_i)-\mathbb{E}[\hat{f}_m(x_i)]\Big)^2\Big] =&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\Big(\sum_{j=1}^{n/m}c_j\mathbf{1}\{x_i\in((j-1)m,jm]\}-\mathbb{E}\big[\sum_{j=1}^{n/m}c_j\mathbf{1}\{x_i\in((j-1)m,jm]\}\big]\Big)^2\Big]\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Bigg[\Big(\sum_{j=1}^{n/m}\big(c_j-\mathbb{E}[c_j]\big)\mathbf{1}\{x_i\in((j-1)m,jm]\}\Big)^2\Bigg]\\
=&\frac{1}{n}\sum_{j=1}^{n/m}\mathbb{E}\Big[\sum_{i=(j-1)m+1}^{jm}\big(c_j-\mathbb{E}[c_j]\big)^2\Big]\\
=&\frac{1}{n}\sum_{j=1}^{n/m}m\mathbb{E}\Big[\big(c_j-\bar{f}^{(j)}\big)^2\Big]\\
=&\frac{1}{n}\sum_{j=1}^{n/m}m\mathbb{E}\Big[\big(\frac{1}{m}\sum_{i=(j-1)m+1}^{jm}y_i-f(x_i)\big)^2\Big]\\
=&\frac{1}{n}\sum_{j=1}^{n/m}m\mathbb{E}\Big[\big(\sum_{i=(j-1)m+1}^{jm}\frac{\epsilon_i}{m}\big)^2\Big]\\
=&\frac{1}{n}\sum_{j=1}^{n/m}\frac{1}{m}\sum_{i=(j-1)m+1}^{jm}\sigma^2\\
=&\frac{\sigma^2}{m}
\end{aligned}
$$

- **Part d:** Derivation of minimal error with respect to m. 1-2 sentences about scaling of m with parameters.

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n/m}\sum_{i=(j-1)m+1}^{jm}\left(\bar{f}^{(j)}-f(x_i)\right)^2 &\leq \frac{L^2}{n^2}(argmin(x_i)-i)^2 \\
&\leq \frac{1}{n}\sum_{j=1}^{n/m}\sum_{i=(j-1)m+1}^{jm}\left(\frac{L}{n}m\right)^2 \\
&= \mathcal{O}\Big(\frac{L^2m^2}{n^2}\Big)
\end{aligned}
$$

Total error is $\mathcal{O}\Big(\frac{L^2m^2}{n^2}+\frac{\sigma^2}{m}\Big)$. If we want to minimize with respect to m, we can set its derivative to zero:

$$
\begin{aligned}
\frac{d}{dm}\Big(\frac{L^2m^2}{n^2}+\frac{\sigma^2}{m}\Big) &= 0 \\
\frac{2L^2m}{n^2}-\frac{\sigma^2}{m^2} &= 0
\end{aligned}
$$

Then we have:

$$
\begin{aligned}
m &= \Big(\frac{\sigma^2n^2}{2L^2}\Big)^{1/3} \\
&= \mathcal{O}\left(\Big(\frac{\sigma n}{L}\Big)^{2/3}\right) = \mathcal{O}\left(\Big(\frac{n}{L}\Big)^{2/3}\right)
\end{aligned}
$$

Plug m back in:

$$
\begin{aligned}
\frac{L^2\big(\frac{n^2\sigma^2}{2L^2}\big)^{2/3}}{n^2}+\sigma^2\Big(\frac{n^2\sigma^2}{2L^2}\Big)^{-1/3} &= \frac{3}{4^{\frac{1}{3}}}\sigma^{\frac{4}{3}}(\frac{L}{n})^{\frac{2}{3}} \\
&= \mathcal{O}\Big((\frac{L}{n})^{\frac{2}{3}}\Big)
\end{aligned}
$$

It just as intuition, minimized total error increased with the decrease of number of samples.