CS 4780/5780 Homework 6 Solution

Problem 1: Regularization Mitigates Overfitting

(a) Due to regularization, $||(\mathbf{w}(\mathcal{D})||_2 \leq B$ for all $\mathcal{D}$. Since $\overline{w} = \mathbb{E}_{\mathcal{D}}(\mathbf{w}(\mathcal{D}))$, we also have,

$$||\overline{w}||_2^2 \leq B^2.$$

Using the triangular inequality we have,

$$||\mathbf{w}(\mathcal{D}) - \overline{w}||_2 \leq ||\mathbf{w}(\mathcal{D})||_2 + ||\overline{w}||_2.$$

Taking the square of each side and asserting inequalities on individual norms,

$$\begin{aligned} ||\mathbf{w}(\mathcal{D}) - \overline{w}||_2^2 &\leq (||\mathbf{w}(\mathcal{D})||_2 + ||\overline{w}||_2)^2 \\ &= ||\mathbf{w}(\mathcal{D})||_2^2 + ||\overline{w}||_2^2 + 2||\mathbf{w}(\mathcal{D})||_2||\overline{w}||_2 \\ &\leq B^2 + B^2 + 2B^2 = 4B^2 \end{aligned}$$

(b) First, note that in $\overline{h}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}(\mathbf{w}(\mathcal{D})^T\mathbf{x})$, the expectation of $\mathbf{w}(\mathcal{D})$ does not depend on $\mathbf{x}$. So

$$\mathbb{E}_{\mathcal{D}}(\mathbf{w}(\mathcal{D})^T\mathbf{x}) = \mathbb{E}_{\mathcal{D}}(\mathbf{w}(\mathcal{D}))^T\mathbf{x} = \overline{w}^T\mathbf{x}$$

We rewrite $h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x})$ as the following:

$$\begin{aligned} h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}) &= \mathbf{w}(\mathcal{D})^T\mathbf{x} - \mathbb{E}_{\mathcal{D}}(\mathbf{w}(\mathcal{D})^T\mathbf{x}) \\ &= \mathbf{w}(\mathcal{D})^T\mathbf{x} - \overline{w}^T\mathbf{x} \\ &= (\mathbf{w}(\mathcal{D}) - \overline{w})^T\mathbf{x} \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$(h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2 \leq \big((\mathbf{w}(\mathcal{D}) - \overline{w})^T\big(\mathbf{w}(\mathcal{D}) - \overline{w}\big)\big)(\mathbf{x}^T\mathbf{x})$$

We write the above in terms of norms and substitute $||\mathbf{x}||_2^2 = 1$.

$$(h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2 \leq ||\mathbf{w}(\mathcal{D}) - \overline{w}||_2^2 \cdot ||\mathbf{x}||_2^2 = ||\mathbf{w}(\mathcal{D}) - \overline{w}||_2^2$$

Using our result from 1a, we get,

$$(h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2 \leq 4B^2$$

Finally, taking the expectation we get,

$$\mathbb{E}_{\mathbf{x},\mathcal{D}}((h_{\mathcal{D}}(\mathbf{x}) - \overline{h}(\mathbf{x}))^2) \leq 4B^2$$

Problem 2: Bias and Variance in KNN

$$\text{EPE}_k(x) = \frac{\sigma^2}{k} + \sigma^2 + \left[f(x) - \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right]^2.$$

where the terms correspond to variance, noise, and bias, respectively.
Indeed, the variance term $\frac{\sigma^2}{k}$ will drop off as k increases.

Derivation:

the error decomposition is, as from lecture,

$$\underbrace{E_{D,(x,y)}\left[(y - h_k(x))^2\right]}_{\text{Expected Test Error}} = \underbrace{E_{x,D}\left[\left(h_k(x) - \overline{h}(x)\right)^2\right]}_{\text{Variance}} + \underbrace{E_{x,y}\left[(\overline{y}(x) - y)^2\right]}_{\text{Noise}} + \underbrace{E_x\left[\left(\overline{h}(x) - \overline{y}(x)\right)^2\right]}_{\text{Bias}}$$

Reminder: $x, x_i, f(x)$, and $f(x_i)$ are treated as constants. $E[\varepsilon] = E[\varepsilon_{(l)}] = 0$. $Var[\varepsilon] = Var[\varepsilon_{(l)}] = \sigma^2$.

$$\overline{y}(x) = E_y\left[y(x)\right] = E\left[f(x) + \varepsilon\right] = f(x) + E\left[\varepsilon\right] = f(x)$$

$$\begin{aligned}
\overline{h}(x) &= E\left[h_k(x)\right] \\
&= E\left[\frac{1}{k}\sum_{l=1}^{k}(f(x_{(l)}) + \varepsilon_{(l)})\right] = \left(\frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right) + E\left[\frac{1}{k}\sum_{l=1}^{k}\varepsilon_{(l)}\right] \\
&= \left(\frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right) + \frac{1}{k}\sum_{l=1}^{k} E\left[\varepsilon_{(l)}\right] = \frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})
\end{aligned}$$

Now we can find the variance, noise, and bias of this classifier.

- Noise:

  $$E_{x,y}\left[(\overline{y}(x) - y)^2\right] = E\left[(f(x) - (f(x) + \varepsilon))^2\right] = E\left[\varepsilon^2\right]$$

  Using the definition of variance,

  $$E\left[\varepsilon^2\right] = E\left[\varepsilon^2\right] - 0^2 = E\left[\varepsilon^2\right] - \left[E\left[\varepsilon\right]\right]^2 = Var[\varepsilon] = \sigma^2$$

- Variance:

  $$E_{x,D}\left[\left(h_k(x) - \overline{h}(x)\right)^2\right] = E\left[\left(\frac{1}{k}\sum_{l=1}^{k}\varepsilon_{(l)}\right)^2\right] = \frac{1}{k^2}E\left[\sum_{1\le i,j\le k}\varepsilon_{(i)}\varepsilon_{(j)}\right] = \frac{1}{k^2}\sum_{1\le i,j\le k} E\left[\varepsilon_{(i)}\varepsilon_{(j)}\right]$$

  For $i \neq j$, the variables ε_i and ε_j are i.i.d.—therefore

  $$E\left[\varepsilon_i\varepsilon_j\right] = E\left[\varepsilon_i\right]E\left[\varepsilon_j\right] = 0$$

  The cross terms in the sum cancel out. We also substitute $E\left[\varepsilon_{(l)}^2\right] = Var[\varepsilon_{(l)}] = \sigma^2$

  $$\frac{1}{k^2}\sum_{1\le i,j\le k} E\left[\varepsilon_{(i)}\varepsilon_{(j)}\right] = \frac{1}{k^2}\sum_{l=1}^{k} E\left[\varepsilon_{(l)}^2\right] = \frac{1}{k^2}\sum_{l=1}^{k}\sigma^2 = \frac{1}{k^2}\cdot k\sigma^2 = \frac{\sigma^2}{k}$$

- Bias:
  Both terms in the expectation are constants given x and D, which are also constants.

  $$E_x\left[\left(\overline{h}(x) - \overline{y}(x)\right)^2\right] = \left(\overline{h}(x) - \overline{y}(x)\right)^2 = \left(\left(\frac{1}{k}\sum_{l=1}^{k} f(x_{(l)})\right) - f(x)\right)^2$$