Homework 8

Runmin Lu

October 30, 2021

Exercise 4.3

(a)

Deterministic noise will go up because f becomes more complex for $\mathcal{H}$ to model.

There's a higher tendency to overfit because overfitting increases as target comlexity increases.

(b)

Deterministic noise will go up because h^* from a simpler $\mathcal{H}$ performs at most as well as h^* from a more complex $\mathcal{H}$ since the simpler $\mathcal{H}$ is a subset of the more complex $\mathcal{H}$.

There's a lower tendency to overfit because a simpler $\mathcal{H}$ is less likely to be led astray by noise.

Exercise 4.5

(a)

Guess $\Gamma = I$

Verify:

$$\begin{aligned}
\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} &= \mathbf{w}^T I^T I\mathbf{w} \\
&= \mathbf{w}^T II\mathbf{w} \\
&= \mathbf{w}^T\mathbf{w} \\
&= \sum_{q=0}^{Q} w_q^2
\end{aligned}$$

(b)

$$\begin{aligned}
\mathbf{w}^T\Gamma^T\Gamma\mathbf{w} &= (\sum_{q=0}^{Q} w_q)^2 \\
&= \sum_{i=0}^{Q} w_i \sum_{j=0}^{Q} w_j \\
&= (\sum_{j=0}^{Q} w_j, ..., \sum_{j=0}^{Q} w_j)\mathbf{w} \\
\mathbf{w}^T\Gamma^T\Gamma &= (\sum_{j=0}^{Q} w_j, ..., \sum_{j=0}^{Q} w_j) \\
\forall \text{ column index } i : \mathbf{w}^T(\Gamma^T\Gamma)_i &= \sum_{j=0}^{Q} w_j \\
&= \mathbf{w}^T\mathbf{1} \text{ where } \mathbf{1} \text{ is the vector of all 1's} \\
\Gamma^T\Gamma &= \begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ ... & ... & ... & ... \\ 1 & 1 & ... & 1 \end{pmatrix} \text{ of dimeision } (Q+1) \times (Q+1)
\end{aligned}$$

The inner product of any 2 columns of Γ is 1.

$$\Gamma = \begin{pmatrix} \frac{1}{\sqrt{Q+1}} & \frac{1}{\sqrt{Q+1}} & ... & \frac{1}{\sqrt{Q+1}} \\ \frac{1}{\sqrt{Q+1}} & \frac{1}{\sqrt{Q+1}} & ... & \frac{1}{\sqrt{Q+1}} \\ ... & ... & ... & ... \\ \frac{1}{\sqrt{Q+1}} & \frac{1}{\sqrt{Q+1}} & ... & \frac{1}{\sqrt{Q+1}} \end{pmatrix}$$

Exercise 4.6

The hard-order constraint is more useful because as the hint says, the norm of $\mathbf{w}$ is irrelevant. If we use the soft-order constraint, then for any $\mathbf{w}$ with $\mathbf{w}^T\mathbf{w} > C$, we can just multiply $\mathbf{w}$ by some positive number α that's small enough to satisfy the constraint but the still perform the same classification because $sign(\mathbf{w}^T\mathbf{x}) = sign(\alpha\mathbf{w}^T\mathbf{x})$ for all $\mathbf{x}$.

Exercise 4.7

(a)

Given that the validation error $E_{\text{val}}(g^-) = \frac{1}{K}\sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{val}}} e(g^-(\mathbf{x}), y)$, we have

$$
\begin{aligned}
\sigma^2_{\text{val}} &= \text{Var}_{\mathcal{D}_{\text{val}}}[E_{\text{val}}(g^-)] \\
&= \text{Var}_{\mathcal{D}_{\text{val}}}[\frac{1}{K}\sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{val}}} e(g^-(\mathbf{x}), y)] \\
&= \frac{1}{K^2}\text{Var}_{\mathcal{D}_{\text{val}}}[\sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{val}}} e(g^-(\mathbf{x}), y)] \\
&= \frac{1}{K^2}\sum_{i=1}^{K}\text{Var}_{\mathbf{x}}[e(g^-(\mathbf{x}), y] \qquad (\mathbf{x}, y) \in \mathcal{D} \text{ are IID} \\
&= \frac{1}{K}\text{Var}_{\mathbf{x}}[e(g^-(\mathbf{x}), y] \\
&= \frac{1}{K}\sigma^2(g^-)
\end{aligned}
$$

(b)

$$
\begin{aligned}
\sigma^2_{\text{val}} &= \frac{1}{K}\text{Var}_{\mathbf{x}}[e(g^-(\mathbf{x}), y)] \\
&= \frac{1}{K}\text{Var}_{\mathbf{x}}[[g^-(\mathbf{x}) \neq y]] \\
&= \boxed{\frac{1}{K}\mathbb{P}[g^-(\mathbf{x}) \neq y](1 - \mathbb{P}[g^-(\mathbf{x}) \neq y])}
\end{aligned}
$$

(c)

Lemma: $f(x) = x(1-x)$ has global maximum $\frac{1}{4}$ Proof:

$$
\begin{aligned}
\frac{df}{dx} &= -2x + 1 = 0 \\
dxx &= \frac{1}{2} \text{ is a critical point} \\
\frac{d^2f}{dx^2} &= -2 < 0 \implies \text{ global max}
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2_{\text{val}} &= \frac{1}{K}\mathbb{P}[g^-(\mathbf{x}) \neq y](1 - \mathbb{P}[g^-(\mathbf{x}) \neq y]) \\
&\leq \frac{1}{K}(\frac{1}{4}) \\
&= \frac{1}{4K}
\end{aligned}
$$

(d)

As the hint says, the squared error is unbounded. y can be as far from $g^{-}(\mathbf{x})$ as possible so there's no uniform upper bound.

(e)

Training with fewer points results in a higher error bar, which results in a higher upper bound for $E_{\text{out}(g^-)}$, which is the mean of $e(g^{-}(\mathbf{x}), y)$. As the hint says, the variance also increases.

(f)

Given that $E_{\text{out}} \leq E_{\text{val}} + O(\frac{1}{\sqrt{K}})$. As we increase K, $O(\frac{1}{\sqrt{K}})$ will get smaller but E_{val}, whose expected value is E_{out}, will get larger since we train using fewer points, resulting in a larger error bar.

4.8

Yes because it only depends on the model, the training data, and the testing data. No other intervention is done in computing E_m.