ECE595 / STAT598: Machine Learning I

Lecture 29.1: Bias and Variance - From VC Analysis to Bias-Variance

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

- Lecture 28 Sample and Model Complexity
- Lecture 29 Bias and Variance
- Lecture 30 Overfit

Today's Lecture:

- From VC Analysis to Bias-Variance
  - Generalization Bound
  - Bias-Variance Decomposition
  - Interpreting Bias-Variance
- Example
  - 0-th order vs 1-st order model
  - Trade off

Generalizing the Generalization Bound

Theorem (Generalization Bound)

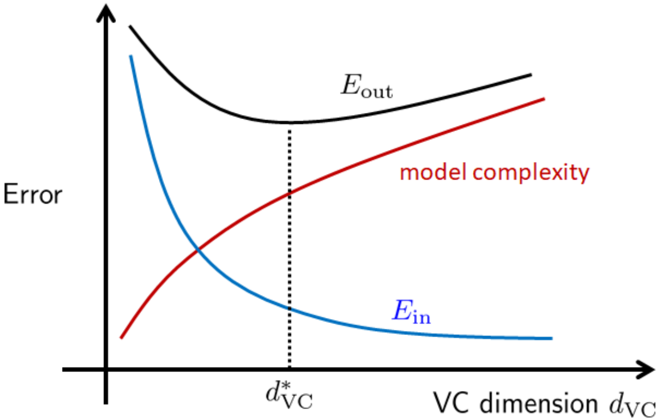
For any tolerance $\delta > 0$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{\mathbf{8}}{N} \log \frac{\mathbf{4} m_{\mathcal{H}}(\mathbf{2N})}{\delta}},$$

with probability at least $1 - \delta$.

- g: final hypothesis
- $m_{\mathcal{H}}(N)$: how complex is your model
- d_{VC}: parameter defining $m_{\mathcal{H}}(N) \leq N^{d_{\text{VC}}} + 1$
- Large d_{VC} $=$ more complex
- So more difficult to train, and hence require more training samples

Trade-off Curve



VC Analysis

- VC analysis is a **decomposition**.
- Decompose E_{out} into E_{in} and ϵ.

$$E_{\text{out}} \leq E_{\text{in}} + \underbrace{\sqrt{\frac{8}{N} \log \frac{4\left((2N)^{d_{\text{VC}}} + 1\right)}{\delta}}}_{=\epsilon}$$

- E_{in} = training error, ϵ = penalty of complex model.
- Bias and variance is another decomposition.
- Decompose E_{out} into
  - How well can $\mathcal{H}$ approximate f?
  - How well can we zoom in a good h in $\mathcal{H}$?
- Roughly speaking we will have

$$E_{\text{out}} = \text{bias} + \textit{variance}$$

From VC Analysis to Bias-Variance

- In **VC analysis** we define the out-sample error as

$$E_{\text{out}}(g) = \mathbb{P}[g(\boldsymbol{x}) \neq f(\boldsymbol{x})]$$

- Let $B = \{g(\boldsymbol{x}) \neq f(\boldsymbol{x})\}$ be the bad event. $B \in \{0, 1\}$.
- Then this is equal to

$$\begin{aligned} E_{\text{out}}(g) &= \mathbb{P}[B = 1] \\ &= 1 \cdot \mathbb{P}[B = 1] + 0 \cdot \mathbb{P}[B = 0] \\ &= \mathbb{E}[B]. \end{aligned}$$

- So $E_{\text{out}}(g)$ can be written as

$$E_{\text{out}}(g) = \mathbb{E}_{\boldsymbol{x}}[\mathbf{1}\{g(\boldsymbol{x}) \neq f(\boldsymbol{x})\}].$$

- Expectation taken over all $\boldsymbol{x} \sim p(\boldsymbol{x})$.

Changing the Error Measure

- In **VC analysis** we define the out-sample error as

$$E_{\text{out}}(g) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbf{1}\{g(\boldsymbol{x}) \neq f(\boldsymbol{x})\}\right]$$

- Expectation of a 0-1 loss.
- In **Bias-variance** analysis we define the out-sample error as

$$E_{\text{out}}(g) = \mathbb{E}_{\boldsymbol{x}}\left[(g(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

- Expectation of a square loss.
- Square loss is differentiable.

Dependency on Training Set

- In VC analysis we define the out-sample error as

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\boldsymbol{x}}\left[\mathbf{1}\{g^{(\mathcal{D})}(\boldsymbol{x}) \neq f(\boldsymbol{x})\}\right]$$

- The final hypothesis depends on $\mathcal{D}$.
- If you use a different $\mathcal{D}$, your g will be different.
- In Bias-variance analysis we define the out-sample error as

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right].$$

- Why did we skip $\mathcal{D}$ in VC analysis?
    - Hoeffding bound is uniform for **all** $\mathcal{D}$
    - So it does not matter which $\mathcal{D}$ you used to generate g
    - Not true for bias-variance

Averaging over all $\mathcal{D}$

- To account for all the possible $\mathcal{D}$'s, compute the expectation and define the expected out-sample error.

$$\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right].$$

- $E_{\text{out}}(g^{(\mathcal{D})})$: Out-sample error for the particular g found from $\mathcal{D}$
- $\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right]$: Out-sample error averaged over all possible $\mathcal{D}$'s
- VC trade-off is a "worst case" analysis
  - Uniform bound on every $\mathcal{D}$
- Bias-variance trade-off is an "average" analysis
  - Average over different $\mathcal{D}$'s

Decomposing $\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})$

- To account for all the possible $\mathcal{D}$'s, compute the expectation and define the expected out-sample error.

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right].$$

- Let us do some calculation

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2 - 2g^{(\mathcal{D})}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\underbrace{\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]}_{\overline{g}(\boldsymbol{x})}f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right].
\end{aligned}$$

The Average $\overline{g}(\boldsymbol{x})$

- The decomposition gives

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{x}}\left[(g^{(\mathcal{D})}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\underbrace{\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]}_{\overline{g}(\boldsymbol{x})} f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right]
\end{aligned}
$$

- We define the term

$$\overline{g}(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]$$

- The asymptotic limit of the estimate

$$\overline{g}(\boldsymbol{x}) \approx \frac{1}{K}\sum_{k=1}^{K} g^{(\mathcal{D}_k)}(\boldsymbol{x})$$

- $g^{(\mathcal{D}_k)}$ are inside the hypothesis set. But $\overline{g}$ is *not* necessarily inside.

Bias and Variance

- Do some additional calculation

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] \\
=\ &\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\boldsymbol{x})]f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] \\
=\ &\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] \\
=\ &\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2 + \overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2\right] \\
=\ &\mathbb{E}_{\boldsymbol{x}}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Big].
\end{aligned}
$$

- Define two terms

$$
\text{bias}(\boldsymbol{x}) \overset{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2,
$$

$$
\text{var}(\boldsymbol{x}) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].
$$

Bias and Variance

- The decomposition:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] \\
= \;& \mathbb{E}_{\boldsymbol{x}}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Big].
\end{aligned}
$$

- Define two terms

$$
\begin{aligned}
\text{bias}(\boldsymbol{x}) &\overset{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2, \\
\text{var}(\boldsymbol{x}) &\overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].
\end{aligned}
$$

- Take expectation

$$
\begin{aligned}
\text{bias} &= \mathbb{E}_{\boldsymbol{x}}[\text{bias}(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x}}\left[(\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right], \\
\text{var} &= \mathbb{E}_{\boldsymbol{x}}[\text{var}(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2]\right].
\end{aligned}
$$

Bias and Variance Decomposition

- The decomposition:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right]$$

$$= \mathbb{E}_{\boldsymbol{x}}\Bigg[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Bigg].$$

- This gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\boldsymbol{x}}[\text{bias}(\boldsymbol{x}) + \text{var}(\boldsymbol{x})] \\ &= \text{bias} + \text{var}\end{aligned}$$

Interpreting the Bias-Variance

- The decomposition:

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\text{out}}(g^{(\mathcal{D})})\right]$$

$$= \mathbb{E}_{\boldsymbol{x}}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\boldsymbol{x})^2\right] - \overline{g}(\boldsymbol{x})^2}_{\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x})-\overline{g}(\boldsymbol{x}))^2]} + \underbrace{\overline{g}(\boldsymbol{x})^2 - 2\overline{g}(\boldsymbol{x})f(\boldsymbol{x}) + f(\boldsymbol{x})^2}_{(\overline{g}(\boldsymbol{x})-f(\boldsymbol{x}))^2}\Big].$$

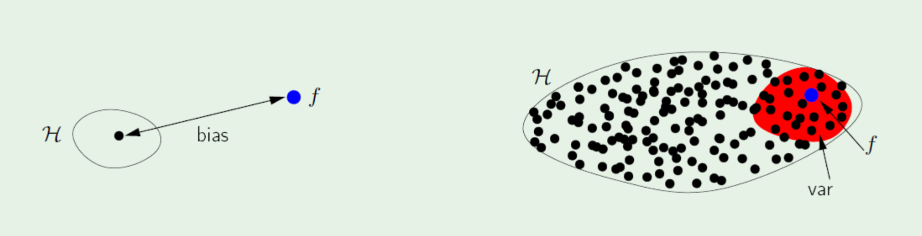
- The two terms:

$$\text{bias}(\boldsymbol{x}) \overset{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2,$$

$$\text{var}(\boldsymbol{x}) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].$$

- bias($\boldsymbol{x}$): How close is the **average function** $\overline{g}$ to the target
- var($\boldsymbol{x}$): How much **uncertainty** you have around $\overline{g}$

Model Complexity



- The bias and variance are

$$\text{bias}(\boldsymbol{x}) \stackrel{\text{def}}{=} (\overline{g}(\boldsymbol{x}) - f(\boldsymbol{x}))^2,$$

$$\text{var}(\boldsymbol{x}) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\boldsymbol{x}) - \overline{g}(\boldsymbol{x}))^2].$$

- If you have a simple $\mathcal{H}$, then large bias but small variance
- If you have a complex $\mathcal{H}$, then small bias but large variance