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 $\delta > 0$

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

with probability at least $1 - \delta$.

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

![Diagram showing the trade-off curve between error and VC dimension. The curve illustrates the relationship between $E_{out}$ and $E_{in}$ as a function of $d_{VC}$. The optimal VC dimension $d^*_C$ is indicated where the error $E_{out}$ and $E_{in}$ are minimized.](image)
VC Analysis

- VC analysis is a **decomposition**.
- Decompose $E_{\text{out}}$ into $E_{\text{in}}$ and $\epsilon$.

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

- $E_{\text{in}} =$ training error, $\epsilon =$ penalty of complex model.
- Bias and variance is another decomposition.
- Decompose $E_{\text{out}}$ 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

$$E_{\text{out}} = \text{bias} + \text{variance}$$
In **VC analysis** we define the out-sample error as

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

Let \( B = \{g(x) \neq f(x)\} \) be the bad event. \( B \in \{0, 1\} \).

Then this is equal to

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

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

\[ E_{\text{out}}(g) = \mathbb{E}_x[1\{g(x) \neq f(x)\}]. \]

Expectation taken over all \( x \sim p(x) \).

Changing the Error Measure

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

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

  Expectation of a **0-1 loss**.

- In **Bias-variance** analysis we define the out-sample error as

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

  Expectation of a **square loss**.

  Square loss is differentiable.
Dependency on Training Set

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

\[ E_{\text{out}}(g^{(D)}) = \mathbb{E}_x \left[ 1\{g^{(D)}(x) \neq f(x)\} \right] \]

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

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

- Why did we skip \( D \) in VC analysis?
  - Hoeffding bound is uniform for all \( D \)
  - So it does not matter which \( D \) 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.

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

- $E_{\text{out}}(g^{(\mathcal{D})})$: 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 $\mathbb{E}_{\text{out}}(g^{(D)})$

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

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

- Let us do some calculation

$$\mathbb{E}_{D} \left[ \mathbb{E}_{x} \left[ (g^{(D)}(x) - f(x))^2 \right] \right]$$

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

$$= \mathbb{E}_{x} \left[ \mathbb{E}_{D} \left[ g^{(D)}(x)^2 - 2g^{(D)}(x)f(x) + f(x)^2 \right] \right]$$

$$= \mathbb{E}_{x} \left[ \mathbb{E}_{D} \left[ g^{(D)}(x)^2 \right] - 2\mathbb{E}_{D}[g^{(D)}(x)f(x)] + f(x)^2 \right].$$
The Average $\bar{g}(x)$

- The decomposition gives
  \[
  \mathbb{E}_D \left[ \mathbb{E}_x \left[ (g^{(D)}(x) - f(x))^2 \right] \right] \\
  = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - 2 \mathbb{E}_D[ g^{(D)}(x) ] f(x) + f(x)^2 \right] 
  \]

- We define the term
  \[
  \bar{g}(x) = \mathbb{E}_D[ g^{(D)}(x) ] 
  \]

- The asymptotic limit of the estimate
  \[
  \bar{g}(x) \approx \frac{1}{K} \sum_{k=1}^{K} g^{(D_k)}(x) 
  \]

- $g^{(D_k)}$ are inside the hypothesis set. But $\bar{g}$ is not necessarily inside.
Bias and Variance

- Do some additional calculation
  \[
  \mathbb{E}_D \left[ \mathbb{E}_{\text{out}}(g^{(D)}) \right]
  = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - 2 \mathbb{E}_D [g^{(D)}(x)] f(x) + f(x)^2 \right]
  = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - 2 \bar{g}(x) f(x) + f(x)^2 \right]
  = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - \bar{g}(x)^2 + \bar{g}(x)^2 - 2 \bar{g}(x) f(x) + f(x)^2 \right]
  = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - \bar{g}(x)^2 + \bar{g}(x)^2 - 2 \bar{g}(x) f(x) + f(x)^2 \right] .
  \]

- Define two terms
  \[
  \text{bias}(x) \overset{\text{def}}{=} (\bar{g}(x) - f(x))^2 ,
  \]
  \[
  \text{var}(x) \overset{\text{def}}{=} \mathbb{E}_D [(g^{(D)}(x) - \bar{g}(x))^2] .
  \]
Bias and Variance

The decomposition:

\[
\mathbb{E}_D \left[ \mathbb{E}_{\text{out}} (g^{(D)}) \right] = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - \bar{g}(x)^2 + \bar{g}(x)^2 - 2\bar{g}(x)f(x) + f(x)^2 \right].
\]

Define two terms

\[
\text{bias}(x) \overset{\text{def}}{=} (\bar{g}(x) - f(x))^2,
\]
\[
\text{var}(x) \overset{\text{def}}{=} \mathbb{E}_D [(g^{(D)}(x) - \bar{g}(x))^2].
\]

Take expectation

\[
\text{bias} = \mathbb{E}_x [\text{bias}(x)] = \mathbb{E}_x \left[ (\bar{g}(x) - f(x))^2 \right],
\]
\[
\text{var} = \mathbb{E}_x [\text{var}(x)] = \mathbb{E}_x \left[ \mathbb{E}_D [(g^{(D)}(x) - \bar{g}(x))^2] \right].
\]
Bias and Variance Decomposition

The decomposition:

\[
\mathbb{E}_D \left[ \mathbb{E}_{\text{out}}(g^{(D)}) \right] \\
= \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 - \overline{g}(x)^2 + \overline{g}(x)^2 - 2\overline{g}(x)f(x) + f(x)^2 \right] \right].
\]

This gives

\[
\mathbb{E}_D \left[ \mathbb{E}_{\text{out}}(g^{(D)}) \right] = \mathbb{E}_x [\text{bias}(x) + \text{var}(x)] \\
= \text{bias} + \text{var}
\]
Interpreting the Bias-Variance

- The decomposition:

\[
\mathbb{E}_D \left[ \mathbb{E}_{\text{out}}(g^{(D)}) \right] = \mathbb{E}_x \left[ \mathbb{E}_D \left[ g^{(D)}(x)^2 \right] - \overline{g(x)}^2 + \overline{g(x)}^2 - 2\overline{g(x)}f(x) + f(x)^2 \right].
\]

- The two terms:

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

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

- bias(x): How close is the \textbf{average function} \(\overline{g}\) to the target
- var(x): How much \textbf{uncertainty} you have around \(\overline{g}\)
The bias and variance are

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

$$\text{var}(x) \overset{\text{def}}{=} \mathbb{E}_D[(g(D)(x) - \bar{g}(x))^2].$$

- 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