CSC 4020 Fundamental of Machine Learning: Bias-Variance Tradeoff

Baoyuan Wu
School of Data Science, CUHK-SZ

January 25/27, 2021

Bias-variance tradeoff

- We are provided by a training dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, which is drawn i.i.d. from some distribution $P(\mathcal{X}, \mathcal{Y})$

Bias-variance tradeoff

- We are provided by a training dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, which is drawn i.i.d. from some distribution $P(\mathcal{X}, \mathcal{Y})$
- The relationship between the input features $\boldsymbol{x}$ and the output y is

$$y = h(\boldsymbol{x}) + e, \ e \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

$$p(y|\boldsymbol{x}) = \mathcal{N}(h(\boldsymbol{x}), \sigma^2 \mathbf{I}), \tag{2}$$

where $h(\boldsymbol{x})$ can be seen as the unknown target function and the mean of $p(y|\boldsymbol{x})$.

Bias-variance tradeoff

- We are provided by a training dataset $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, which is drawn i.i.d. from some distribution $P(\mathcal{X}, \mathcal{Y})$
- The relationship between the input features $\boldsymbol{x}$ and the output y is

$$y = h(\boldsymbol{x}) + e, \ e \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

$$p(y|\boldsymbol{x}) = \mathcal{N}(h(\boldsymbol{x}), \sigma^2 \mathbf{I}), \tag{2}$$

where $h(\boldsymbol{x})$ can be seen as the unknown target function and the mean of $p(y|\boldsymbol{x})$.

- The goal of machine learning is to learn a hypothesis function based on the training dataset D using some learning algorithm $\mathcal{A}$, *i.e.*,

$$h_D = \mathcal{A}(D)$$

Bias-variance tradeoff

- **Expected hypothesis function** (given $\mathcal{A}$):

$$\bar{h} = E_{D \sim P^n}[h_D] = \int_D h_D p(D) dD$$

Bias-variance tradeoff

- **Expected hypothesis function** (given $\mathcal{A}$):

$$\bar{h} = E_{D\sim P^n}[h_D] = \int_D h_D p(D) dD$$

- Given a test pair $(\boldsymbol{x}, y) \sim P(\mathcal{X}, \mathcal{Y})$ and h_D, the **expected test error** is defined as

$$E_{(\boldsymbol{x},y)\sim P}\left[(h_D(\boldsymbol{x}) - y)^2\right] = \int_{\boldsymbol{x}} \int_y (h_D(\boldsymbol{x}) - y)^2 p(x, y) d\boldsymbol{x} dy$$

Bias-variance tradeoff

- **Expected hypothesis function** (given $\mathcal{A}$):

$$\bar{h} = E_{D\sim P^n}[h_D] = \int_D h_D p(D) dD$$

- Given a test pair $(\boldsymbol{x}, y) \sim P(\mathcal{X}, \mathcal{Y})$ and h_D, the **expected test error** is defined as

$$E_{(\boldsymbol{x},y)\sim P}\left[(h_D(\boldsymbol{x}) - y)^2\right] = \int_{\boldsymbol{x}} \int_y (h_D(\boldsymbol{x}) - y)^2 p(x, y) d\boldsymbol{x} dy$$

- Given a test pair $(\boldsymbol{x}, y) \sim P(\mathcal{X}, \mathcal{Y})$ and $\mathcal{A}$, the **expected test error** is defined as

$$E_{(\boldsymbol{x},y)\sim P, D\sim P^n}\left[(h_D(\boldsymbol{x}) - y)^2\right] = \int_D \int_{\boldsymbol{x}} \int_y (h_D(\boldsymbol{x}) - y)^2 p(x, y) d\boldsymbol{x} dy dD$$

Bias-variance tradeoff

- **Expected hypothesis function** (given $\mathcal{A}$):

$$\bar{h} = E_{D\sim P^n}[h_D] = \int_D h_D p(D) dD$$

- Given a test pair $(\boldsymbol{x}, y) \sim P(\mathcal{X}, \mathcal{Y})$ and h_D, the **expected test error** is defined as

$$E_{(\boldsymbol{x},y)\sim P}\left[(h_D(\boldsymbol{x}) - y)^2\right] = \int_{\boldsymbol{x}} \int_y (h_D(\boldsymbol{x}) - y)^2 p(x, y) d\boldsymbol{x} dy$$

- Given a test pair $(\boldsymbol{x}, y) \sim P(\mathcal{X}, \mathcal{Y})$ and $\mathcal{A}$, the **expected test error** is defined as

$$E_{(\boldsymbol{x},y)\sim P, D\sim P^n}\left[(h_D(\boldsymbol{x}) - y)^2\right] = \int_D \int_{\boldsymbol{x}} \int_y (h_D(\boldsymbol{x}) - y)^2 p(x, y) d\boldsymbol{x} dy dD$$

- We are interested in evaluating the quality of a machine learning algorithm $\mathcal{A}$ with respect to a data distribution $P(\mathcal{X}, \mathcal{Y})$. In the following we will show that this expression decomposes into three meaningful terms.

Bias-variance tradeoff

- The **expected test error** can be decomposed as follows

$$
\begin{aligned}
&E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] = E_{(\boldsymbol{x},y),D}\left[[h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x})) + (\bar{h}(\boldsymbol{x}) - y)]^2\right] \\
=&E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + 2E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x}) - y)\right] \\
&+ E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x}) - y)^2\right]
\end{aligned}
$$

Bias-variance tradeoff

- The **expected test error** can be decomposed as follows

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-y)^2\right] = E_{(\boldsymbol{x},y),D}\left[[h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x})) + (\bar{h}(\boldsymbol{x})-y)]^2\right] \\
=& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x}))^2\right] + 2E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x})-y)\right] \\
& + E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x})-y)^2\right]
\end{aligned}
$$

- We have

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x})-y)\right] \\
=& E_{(\boldsymbol{x},y)}\left[E_D[(h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x})-y)]\right] \\
=& E_{(\boldsymbol{x},y)}\left[(E_D[(h_D(\boldsymbol{x})]-\bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x})-y)\right) = 0
\end{aligned}
$$

Bias-variance tradeoff

- The **expected test error** can be decomposed as follows

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] = E_{(\boldsymbol{x},y),D}\left[[h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x})) + (\bar{h}(\boldsymbol{x}) - y)]^2\right] \\
=& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + 2E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x}) - y)\right] \\
& + E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x}) - y)^2\right]
\end{aligned}
$$

- We have

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x}) - y)\right] \\
=& E_{(\boldsymbol{x},y)}\left[E_D[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x}) - y)]\right] \\
=& E_{(\boldsymbol{x},y)}\left[(E_D[(h_D(\boldsymbol{x})] - \bar{h}(\boldsymbol{x}))(\bar{h}(\boldsymbol{x}) - y)\right) = 0
\end{aligned}
$$

- Then, we have

$$
E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] = E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x}) - y)^2\right]
$$

Bias-variance tradeoff

- We also have

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x})-y)^2\right] = E_{(\boldsymbol{x},y),D}\left[[(\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))+(h(x)-y)]^2\right] \\
=& E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)^2\right] + E_{\boldsymbol{x},y}\left[\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))^2\right] + 2E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)(\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))\right] \\
=& E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)^2\right] + E_{\boldsymbol{x},y}\left[\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))^2\right] \qquad (3)
\end{aligned}
$$

Bias-variance tradeoff

- We also have

$$
\begin{aligned}
&E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x})-y)^2\right]=E_{(\boldsymbol{x},y),D}\left[[(\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))+(h(x)-y)]^2\right]\\
=&E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)^2\right]+E_{\boldsymbol{x},y}\left[\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))^2\right]+2E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)(\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))\right]\\
=&E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)^2\right]+E_{\boldsymbol{x},y}\left[\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))^2\right] \qquad (3)
\end{aligned}
$$

- Finally, we have

$$
\begin{aligned}
&E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-y)^2\right]\\
=&E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x}))^2\right]+E_{(\boldsymbol{x},y)}\left[(\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))^2\right]+E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)^2\right]
\end{aligned}
$$

Bias-variance tradeoff

- We also have

$$
\begin{aligned}
&E_{(\boldsymbol{x},y),D}\left[(\bar{h}(\boldsymbol{x}) - y)^2\right] = E_{(\boldsymbol{x},y),D}\left[[(\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x})) + (h(x) - y)]^2\right] \\
=&E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)^2\right] + E_{\boldsymbol{x},y}\left[\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))^2\right] + 2E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)(\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))\right. \\
=&E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)^2\right] + E_{\boldsymbol{x},y}\left[\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))^2\right] \qquad (3)
\end{aligned}
$$

- Finally, we have

$$
\begin{aligned}
&E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] \\
=&E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + E_{(\boldsymbol{x},y)}\left[(\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))^2\right] + E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)^2\right]
\end{aligned}
$$

- Above three terms are **variance, bias, noise**, respectively.

Bias-variance tradeoff

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] \\
=& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + E_{(\boldsymbol{x},y)}\left[(\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))^2\right] + E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)^2\right]
\end{aligned}
$$

Bias-variance tradeoff

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-y)^2\right] \\
=& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x})-\bar{h}(\boldsymbol{x}))^2\right] + E_{(\boldsymbol{x},y)}\left[(\bar{h}(\boldsymbol{x})-h(\boldsymbol{x}))^2\right] + E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x})-y)^2\right]
\end{aligned}
$$

- **variance**: Captures how much your classifier changes if you train on a different training set. How "over-specialized" is your classifier to a particular training set (overfitting)? If we have the best possible model for our training data, how far off are we from the average classifier?

Bias-variance tradeoff

$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] \\
=& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + E_{(\boldsymbol{x},y)}\left[(\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))^2\right] + E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)^2\right]
\end{aligned}
$$

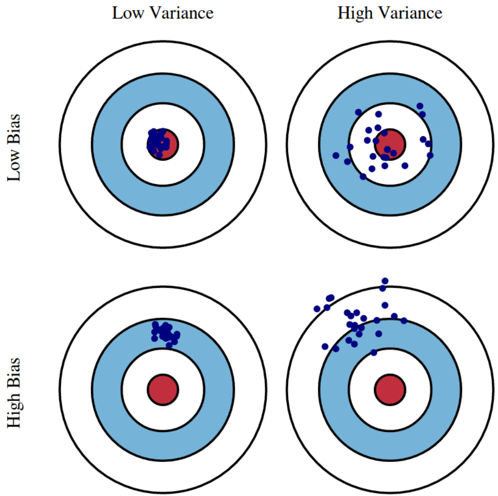
- **variance**: Captures how much your classifier changes if you train on a different training set. How "over-specialized" is your classifier to a particular training set (overfitting)? If we have the best possible model for our training data, how far off are we from the average classifier?
- **Bias**: What is the inherent error that you obtain from your classifier even with infinite training data? This is due to your classifier being "biased" to a particular kind of solution (*e.g.*, linear classifier). In other words, bias is inherent to your model.

Bias-variance tradeoff

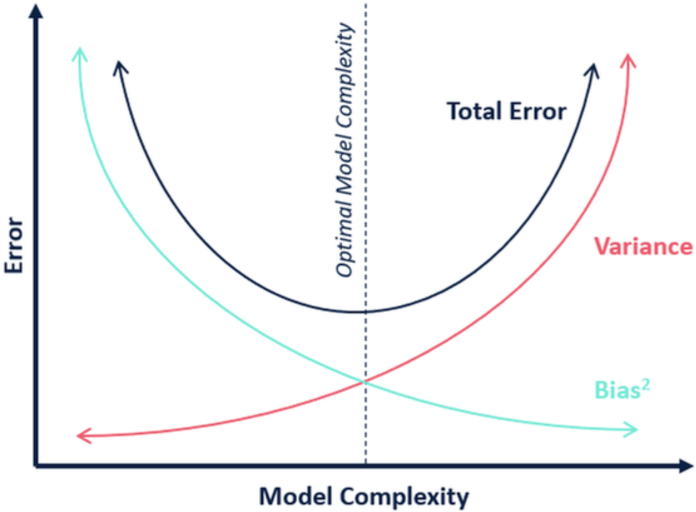
$$
\begin{aligned}
& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - y)^2\right] \\
=& E_{(\boldsymbol{x},y),D}\left[(h_D(\boldsymbol{x}) - \bar{h}(\boldsymbol{x}))^2\right] + E_{(\boldsymbol{x},y)}\left[(\bar{h}(\boldsymbol{x}) - h(\boldsymbol{x}))^2\right] + E_{\boldsymbol{x},y}\left[(h(\boldsymbol{x}) - y)^2\right]
\end{aligned}
$$

- **variance**: Captures how much your classifier changes if you train on a different training set. How "over-specialized" is your classifier to a particular training set (overfitting)? If we have the best possible model for our training data, how far off are we from the average classifier?
- **Bias**: What is the inherent error that you obtain from your classifier even with infinite training data? This is due to your classifier being "biased" to a particular kind of solution (*e.g.*, linear classifier). In other words, bias is inherent to your model.
- **Noise**: How big is the data-intrinsic noise? This error measures ambiguity due to your data distribution and feature representation. You can never beat this, it is an aspect of the data.

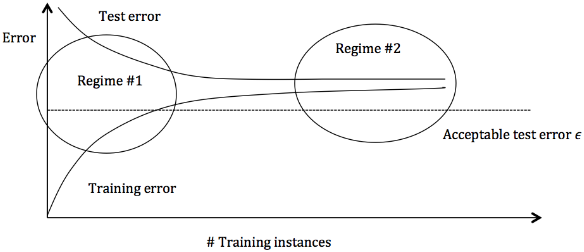
Bias-variance tradeoff



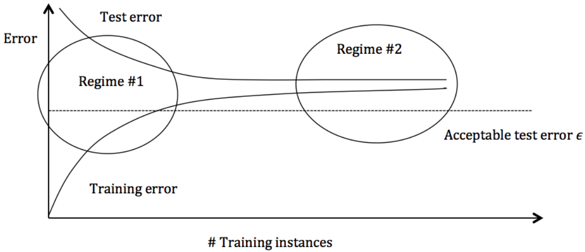
Bias-variance tradeoff



Bias-variance tradeoff



Bias-variance tradeoff



Regime 1 (High Variance)

In the first regime, the cause of the poor performance is high variance.

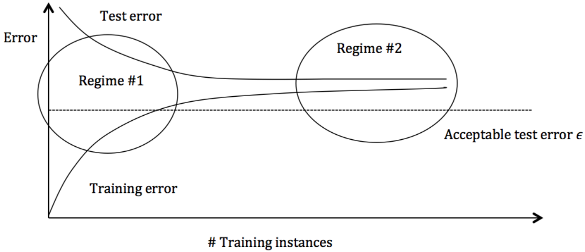
Symptoms:

1. Training error is much lower than test error
2. Training error is lower than ϵ
3. Test error is above ϵ

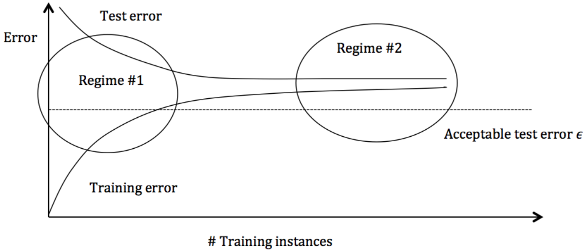
Remedies:

- Add more training data
- Reduce model complexity -- complex models are prone to high variance

Bias-variance tradeoff



Bias-variance tradeoff



Regime 2 (High Bias)

Unlike the first regime, the second regime indicates high bias: the model being used is not robust enough to produce an accurate prediction.

Symptoms:

1. Training error is higher than ϵ

Remedies:

- Use more complex model (e.g. kernelize, use non-linear models)
- Add features

Bias-variance tradeoff

More details can be found at `https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote12.html`