Bias-Variance Tradeoff

CS534 - Machine Learning

Yubin Park, PhD

Complex model *vs.* Simple model

Which way to go?

Bias-Variance Framework can provide

some guidance

Consider a system

$$y = g(\mathbf{x}) + \epsilon$$

where $\mathrm{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] = \sigma^2$

Notations

Assume that we generate a finite number of samples from this system.

$\mathrm{E}[\cdot]$: Expectation operator. e.g. $\mathrm{E}[y]$ is the expected value of y.

$\mathrm{Var}[\cdot]$: Variance operator.

g: True model

f: Our model trained on **finite** samples

$\mathrm{E}[f]$: Expectation of f. In other words, f trained on **infinite** samples.

$\mathrm{Var}[f]$: Variance of f. In other words, the variability of f due to the sampling nature. Note $\mathrm{Var}[f] = \mathrm{E}[f^2] - \mathrm{E}[f]^2$

$\mathrm{E}[(y - f)^2]$: Expected error. NOTE that this is not the same as training or test errors.

$$
\begin{aligned}
\mathrm{E}[(y - f)^2] &= \mathrm{E}[(g + \epsilon - f + \mathrm{E}[f] - \mathrm{E}[f])^2] \\
&= \mathrm{E}[((g - \mathrm{E}[f]) + (\mathrm{E}[f] - f) + \epsilon)^2] \\
&= \mathrm{E}[(g - \mathrm{E}[f])^2] + \mathrm{E}[(\mathrm{E}[f] - f))^2] + \mathrm{E}[\epsilon^2] \\
&= \mathrm{Bias}[f]^2 + \mathrm{Var}[f] + \sigma^2
\end{aligned}
$$

The error consists of

Bias, Variance, and Irreducible Error.

What is Bias?

$$\mathrm{Bias}[f] = g - \mathrm{E}[f]$$

Bias is the difference between the true model, g, and the (theoretical) best-case of our model, $\mathrm{E}[f]$

For example, consider $g(x) = \exp(x)$, while $f(x) = \sum_{k=0}^{K} \beta_k x^k$.

As finite-degree polynomial functions cannot behave like exponential functions, there will be always some gap between the true and our models.

This theoretical gap is called "bias".

Bias is related to the expressibility (or representational capacity) of a model.

In general, simpler models have higher bias.

What is Variance?

$$\mathrm{Var}[f] = \mathrm{E}[(\mathrm{E}[f] - f))^2]$$

Variance is the variability of our model due to the sampling nature of the data.

For example, the coefficients, β, of a linear model will vary depending on selected samples.

The degree of such variability is called the variance.

In other words, our model has high variance if our model changes a lot with inclusion/exlusion of data samples.

NOTE that we are measuring the variability against the expectation of our model, not the true model.

In general, simpler models have less variance.

Conceptually something like this?

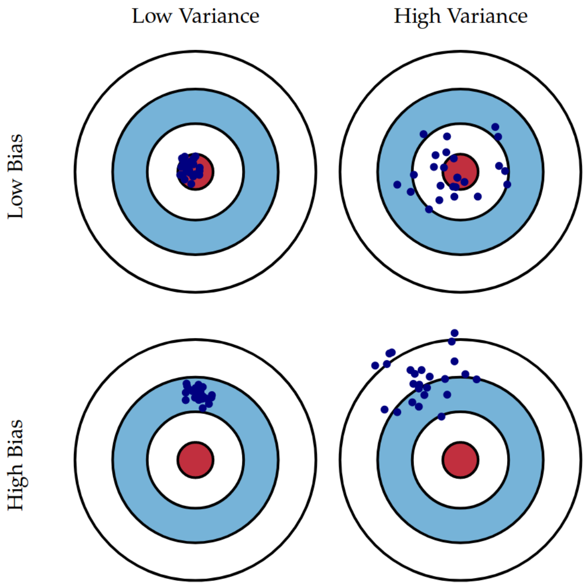


Fig. 1 Graphical illustration of bias and variance.

http://scott.fortmann-roe.com/docs/BiasVariance.html

Model Complexity and Bias-Variance

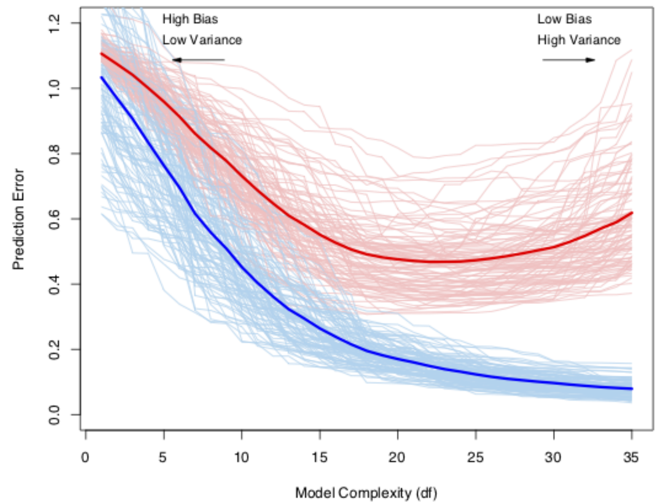


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\mathrm{err}}$, *while the light red curves show the conditional test error* $\mathrm{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\mathrm{E}[\overline{\mathrm{err}}]$.

Chapter 7 of https://web.stanford.edu/~hastie/ElemStatLearn/

Why Shrinkage Models can be Better?

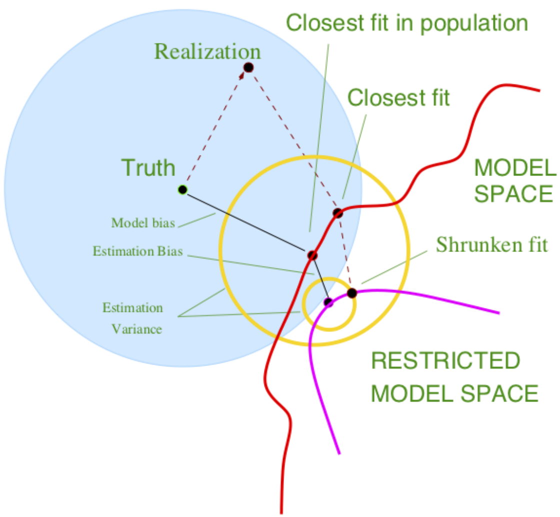
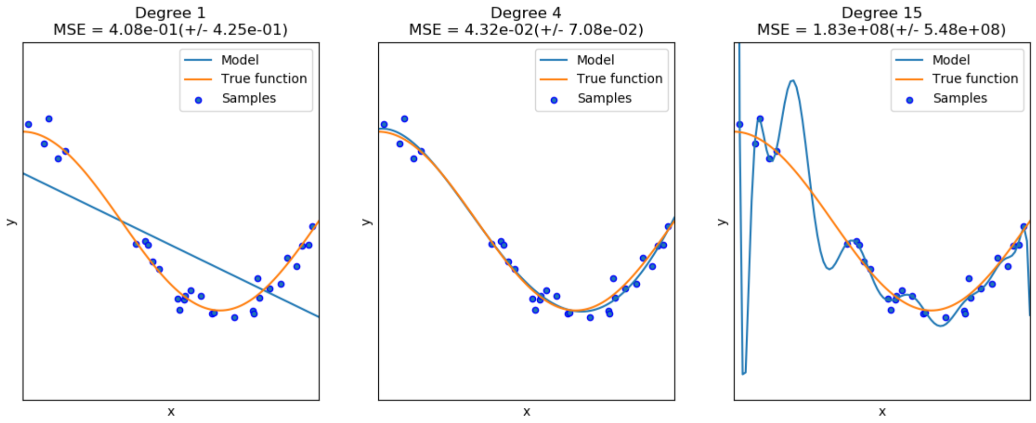


FIGURE 7.2. *Schematic of the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the "closest fit" labeled with a black dot. The model bias from the truth is shown, along with the variance, indicated by the large yellow circle centered at the black dot labeled "closest fit in population." A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.*

Chapter 7 of https://web.stanford.edu/~hastie/ElemStatLearn/

Overfitting vs. Underfitting



Underfitting (left most) = High Bias

Overfitting (right most) = High Variance

https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Assume the Irreducible Error is very small

and the size of training data is fairly large.

$$\mathrm{Bias}^2 = \mathrm{E}[(g - \mathrm{E}[f])^2]$$

$$\approx \mathrm{Average}[(y - f)^2]$$

$$= \mathrm{Training\ Error}$$

Assume our model is complex enough

i.e. $\mathrm{E}[f] = g$

and the Irreducible Error is ver small

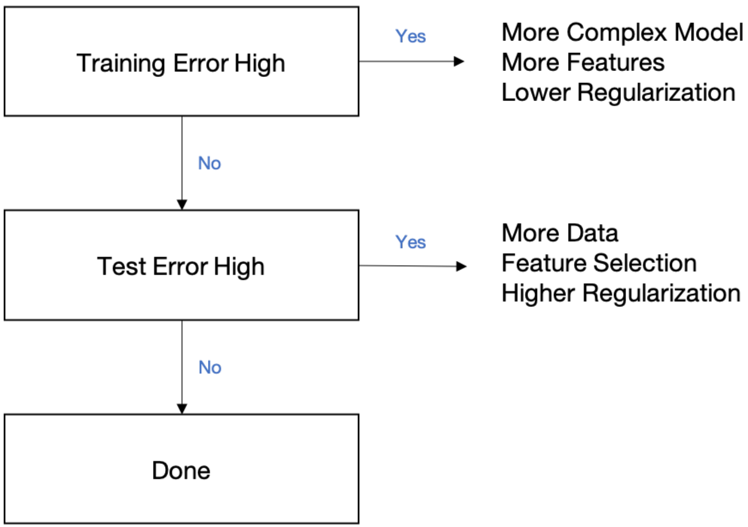
and also the size of test data is fairly large.

$$\text{Variance} = E[(f - \mathrm{E}[f])^2]$$

$$\approx \text{Average}[(y - f)^2]$$

$$= \text{Test Error}$$

How to Use the Bias-Variance



Adapted from Andrew Ng's "Applied Bias-Variance for Deep Learning Flowchart" for building better deep learning systems,
http://www.computervisionblog.com/2016/12/nuts-and-bolts-of-building-deep.html

Questions?