Bias-Variance Trade-off

Intro 2 Statistical Learning

Gaston Sanchez

Bias-Variance Trade-off

Bias-Variance Wikipedia Entry

bias variance tradeoff

All Images Videos News Shopping More Settings Tools

About 374,000 results (0.39 seconds)

Bias–variance tradeoff - Wikipedia

https://en.wikipedia.org/wiki/Bias–variance_tradeoff

In statistics and machine learning, the **bias–variance tradeoff** (or dilemma) is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithms from generalizing beyond their training set: The bias is an error from erroneous assumptions in the learning algorithm. High bias can cause ...

Motivation · Bias–variance ... · Application to classification · Approaches

You should know

The “bias-variance decomposition” is a **conceptual device** based on the (theoretical) expected squared prediction error of an estimated model.

Keep in mind that this decomposition is of a theoretical nature, and its derivation involves assuming a population data set.

Theoretical Considerations

Suppose we could have an infinite number of independent *training* sets of the same size, and we use them to fit an infinite number of models.

Likewise, suppose we had an ideal *test* set, independent from the training sets, from which we draw an observation x_0, and we compute infinite predictions $\hat{f}(x_0)$.

In this idealized situation, errors will still occur because no learning scheme is perfect:

- we have noise in the data (i.e. ϵ)
- not all models will perfectly fit x_0

Theoretical Considerations

A key question of interest is: What is the expected error when predicting x_0? (i.e. *generalization error* on new data?)

Using squared-error loss we have:

$$\text{Err}(x_0) = E[(Y - \hat{f}(x_0))^2 | X = x_0]$$

It can be shown that this expected error can be decomposed in three pieces as:

$$\text{Err}(x_0) = \sigma^2 + \text{Bias}^2(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0))$$

This is the so-called *Bias-Variance Decomposition*

MSE of a Statistical Estimator

To explain where the bias-variance decomposition comes from, we need to review some basic concepts of statistical estimators.

The main concept has to do with the Mean Squared Error of an estimator.

Reminder of Statistical Estimation

Reminder: Estimation

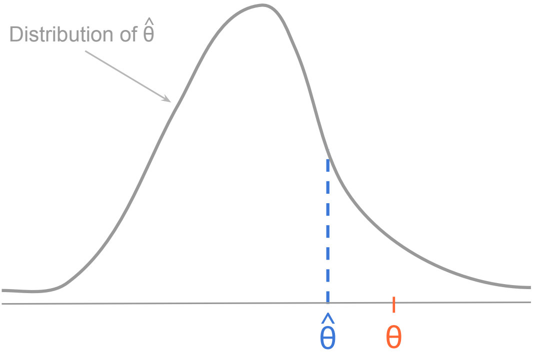
Simply put, estimation consists of providing an approximate value to the parameter of a population, using a (random) sample of observations drawn from such population.

Considerations

- Consider a population for which we want to estimate an unknown value θ (i.e. parameter)
- Say we use an estimator $\hat{\theta}$ based on our sample data.
- Suppose that $\hat{\theta}$ has a finite variance.
- Also, suppose that we know the distribution of $\hat{\theta}$.

Keep in mind

- An estimator is a random variable
- A first sample will result in $\hat{\theta}_1$
- A second sample will result in $\hat{\theta}_2$
- A third sample will result in $\hat{\theta}_3$
- and so on ...
- Some samples will yield a $\hat{\theta}$ that overestimates θ
- Other samples will yield a $\hat{\theta}$ that underestimates θ
- Some samples will yield a $\hat{\theta}$ matching θ



How much different—or similar—is $\hat{\theta}$ from θ?

i.e. how accurate is $\hat{\theta}$?

Estimation Error

A natural question that we can ask is:
How different is $\hat{\theta}$ from θ?

This question involves looking at the difference: $\hat{\theta} - \theta$, which is commonly referred to as the *estimation error*:

$$\text{estimation error} = \hat{\theta} - \theta$$

We would like to measure the "size" of such difference.

Estimation Error

Notice that the estimation error is also a random variable:

- A first sample will result in an error $\hat{\theta}_1 - \theta$
- A second sample will result in an error $\hat{\theta}_2 - \theta$
- A third sample will result in an error $\hat{\theta}_3 - \theta$
- and so on ...

So how do we measure the "size" of the estimation errors?

MSE of a Statistical Estimator

The typical way to quantify the amount of estimation error is by calculating the squared errors, and then averaging over all the possible values of the estimators.

This is known as the **Mean Squared Error** (MSE) of $\hat{\theta}$:

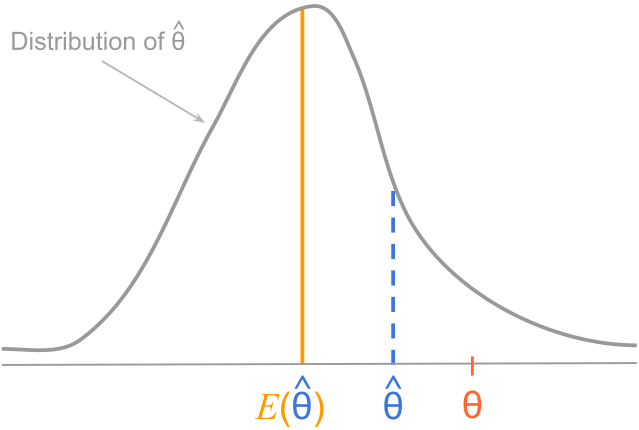
$$\mathsf{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

MSE of a Statistical Estimator

We use the **Mean Squared Error** to measure the accuracy of an estimator $\hat{\theta}$.

MSE is the squared distance from our estimator $\hat{\theta}$ to the true value θ, averaged over all possible samples.

It is convenient to regard the estimation error, $\hat{\theta} - \theta$, with respect to $E(\hat{\theta})$ (see diagram in next slide)


Distribution of $\hat{\theta}$
$E(\hat{\theta})$
$\hat{\theta}$
θ

MSE of a Statistical Estimator

Let's rewrite $(\hat{\theta} - \theta)^2$ as $(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2$, and let $E(\hat{\theta}) = \mu_{\hat{\theta}}$. Then:

$$\begin{aligned}
(\hat{\theta} - \theta)^2 &= \left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right)^2 \\
&= (\hat{\theta} - \mu_{\hat{\theta}} + \mu_{\hat{\theta}} - \theta)^2 \\
&= (\underbrace{\hat{\theta} - \mu_{\hat{\theta}}}_{a} + \underbrace{\mu_{\hat{\theta}} - \theta}_{b})^2 \\
&= a^2 + b^2 + 2ab \\
\implies E\left[(\hat{\theta} - \theta)^2\right] &= E[a^2 + b^2 + 2ab]
\end{aligned}$$

MSE of a Statistical Estimator

We have that $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ can be decomposed as:

$$\begin{aligned} E\left[(\hat{\theta} - \theta)^2\right] &= E[a^2 + b^2 + 2ab] \\ &= E(a^2) + E(b^2) + 2E(ab) \\ &= E[(\hat{\theta} - \mu_{\hat{\theta}})^2] + E[(\mu_{\hat{\theta}} - \theta)^2] + 2E(ab) \end{aligned}$$

Notice that $E(ab)$:

$$E(ab) = E[(\hat{\theta} - \mu_{\hat{\theta}})(\mu_{\hat{\theta}} - \theta)] = 0$$

MSE of a Statistical Estimator

$$\mathsf{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

$$= E[(\hat{\theta} - \mu_{\hat{\theta}})^2] + E[(\mu_{\hat{\theta}} - \theta)^2]$$

$$= E[(\underbrace{\hat{\theta} - \mu_{\hat{\theta}}}_{\mathsf{Bias}})^2] + \underbrace{E[(\mu_{\hat{\theta}} - \theta)^2]}_{\mathsf{Variance}}$$

$$= \mathsf{Bias}^2(\hat{\theta}) + \mathsf{Var}(\hat{\theta})$$

In Summary

The MSE of an estimator can be decomposed in terms of Bias and Variance.

Bias, $\hat{\theta} - \mu_{\hat{\theta}}$, is the tendency of $\hat{\theta}$ to overestimate or underestimate θ over all possible samples.

Variance, $\text{Var}(\hat{\theta})$, simply measures the average variability of the estimators around their mean $E(\hat{\theta})$.

Representative Scenarios for Bias-Variance

Bias-Variance Decomposition for $\hat{f}(X) - f(X)$

In Regression Models ...

In regression models, $\hat{f}(X)$ is an estimator of $f(X)$. Thus, we could ask about the mean squared error of $\hat{f}(X)$

$$\text{MSE}(\hat{f}(X)) = E[(\hat{f}(X) - f(X))^2]$$

In Regression Models ...

In order to improve readibility, let's represent $\hat{f}(x)$ simply as $\hat{f}$:

$$\begin{aligned}
(\hat{f} - f)^2 &= (\hat{f} - E(\hat{f}) + E(\hat{f}) - f)^2 \\
&= (\underbrace{\hat{f} - E(\hat{f})}_{a} + \underbrace{E(\hat{f}) - f}_{b})^2 \\
&= a^2 + b^2 + 2ab \\
\implies E\left[(\hat{f} - f)^2\right] &= E[a^2 + b^2 + 2ab]
\end{aligned}$$

MSE of a Statistical Estimator

We have that MSE$(\hat{f}) = E[(\hat{f} - f)^2]$ can be decomposed as:

$$\begin{aligned} E\left[(\hat{f} - f)^2\right] &= E[a^2 + b^2 + 2ab] \\ &= E(a^2) + E(b^2) + 2E(ab) \\ &= E[(\hat{f} - E(\hat{\theta}))^2] + E[(E(\hat{f}) - f)^2] + 2E(ab) \end{aligned}$$

Notice that $E(ab)$:

$$E(ab) = E[(\hat{f} - E(\hat{f}))(E(\hat{f}) - f)] = 0$$

MSE of a Statistical Estimator

$$\mathsf{MSE}(\hat{f}) = E\left[(\hat{f} - f)^2\right]$$

$$= E[(\hat{f} - E(\hat{f}))^2] + E[(E(\hat{f}) - f)^2]$$

$$= E[(\underbrace{\hat{f} - E(\hat{f})}_{\text{Bias}})^2] + \underbrace{E[(E(\hat{f}) - f)^2]}_{\text{Variance}}$$

$$= \mathsf{Bias}^2(\hat{f}) + \mathsf{Var}(\hat{f})$$

Example

To make things less abstract, let's consider a hypothetical example (stolen from Norman Matloff, 2017)

- Consider a chain of hospitals.
- They are interested in comparing the quality of care for hear attack patients.
- They want to compare the level of quality of care for different locations.
- Let's see how bias and variance may occur in this case. *Example discussed in class.*

Bias-Variance Decomposition for $Y - \hat{f}(X)$

In Regression Models ...

In regression models, we also use $\hat{f}(X)$ to obtain predictions of Y, assuming the standard conceptual equation:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon$$

where:

- $E(\varepsilon) = 0$
- $Var(\varepsilon) = \sigma^2$

In Regression Models ...

Recall that we have two flavors of predictions:

The training set is formed by observations (x_i, y_i) that are used to train the model. And we can calculate a training MSE. This is a measure of resubstition or apparent error.

The test set is formed by observations (x_0, y_0) that are NOT used to train the model. We can calculate a test MSE. This is a measure of generalization or error.

In Regression Models ...

From a theoretical point of view, one population value (i.e. parameter) that we are interested in estimating is the *test MSE*

$$\text{Population Test MSE} = E[(y_0 - \hat{f}(x_0))^2]$$

In Regression Models ...

Let's focus on $(y - \hat{f}(x))^2$. In order to improve readibility, let's represent $\hat{f}(x)$ simply as $\hat{f}$:

$$\begin{aligned}
(y - \hat{f}(x))^2 &= (y - \hat{f})^2 \\
&= (f + \epsilon - \hat{f})^2 \\
&= (f + \epsilon - E(\hat{f}) + E(\hat{f}) - \hat{f})^2 \\
&= (\underbrace{f - E(\hat{f}) + \epsilon}_{a} - \underbrace{[\hat{f} - E(\hat{f})]}_{b})^2 \\
&= a^2 + b^2 - 2ab
\end{aligned}$$

In Regression Models ...

Let's see what's going on with: $(y - \hat{f})^2 = a^2 + b^2 - 2ab$

$$
\begin{aligned}
a^2 &= (f - E(\hat{f}))^2 + \epsilon^2 + 2\epsilon[f - E(\hat{f})] \\
b^2 &= (\hat{f} - E(\hat{f}))^2 \\
2ab &= 2[f - E(\hat{f}) + \epsilon][\hat{f} - E(\hat{f})]
\end{aligned}
$$

But we need to find the expectations:

$$E[(y - \hat{f})^2] = E(a^2) + E(b^2) - 2E(ab)$$

Test MSE

$$E(a^2) = E\left[(f - E(\hat{f}))^2\right] + E(\epsilon^2) + 2E(\epsilon[f - E(\hat{f})])$$
$$E(b^2) = E\left[(\hat{f} - E(\hat{f}))^2\right]$$
$$E(2ab) = 2E\left([f - E(\hat{f}) + \epsilon][\hat{f} - E(\hat{f})]\right)$$

Notice that:

$$E(a^2) = E\left[(f - E(\hat{f}))^2\right] + E(\epsilon^2)$$
$$E(b^2) = E\left[(\hat{f} - E(\hat{f}))^2\right]$$
$$E(2ab) = 0$$

Bias-Variance Decomposition

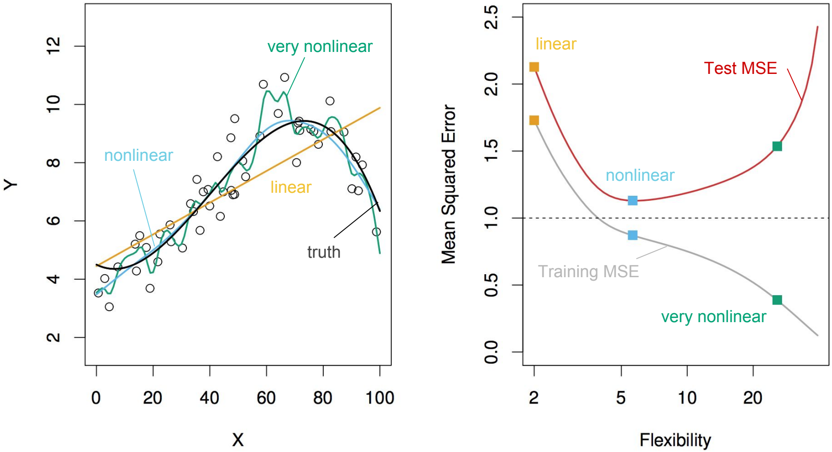
$$
\begin{aligned}
E[(y-\hat{f})^2] &= E\left[(f-E(\hat{f}))^2\right] + E(\epsilon^2) + E\left[(\hat{f}-E(\hat{f}))^2\right] \\
&= E(\epsilon^2) + E\left[(f-E(\hat{f}))^2\right] + E\left[(\hat{f}-E(\hat{f}))^2\right] \\
&= \sigma^2 + \text{Bias}^2(\hat{f}) + \text{Variance}(\hat{f}) \\
&= \text{Irreducible Error} + \text{Bias}^2 + \text{Variance}
\end{aligned}
$$

Bias-Variance Decomposition

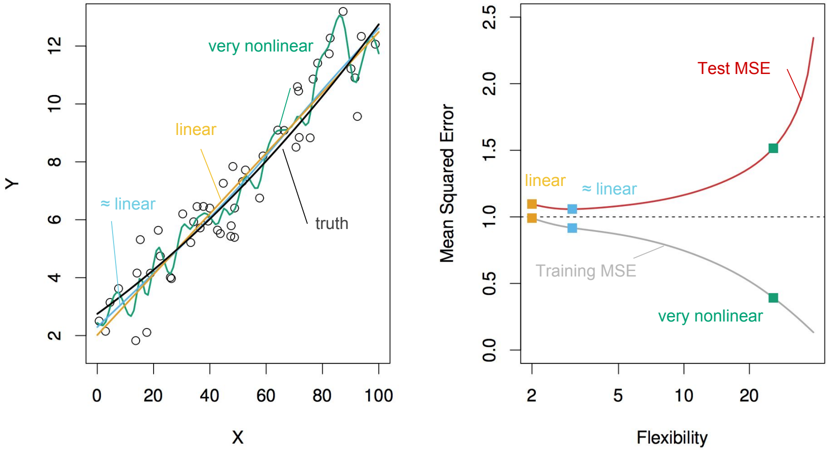
- The irreducible error σ^2 cannot be avoided no matter how well we estimate $f(x)$, unless $\sigma^2 = 0$.
- The squared bias is the amount by which the average of the estimate differs from the true mean.
- The variance is the expected squared deviation of $\hat{f}(x)$ around its mean.
- Typically, the more complex the model $\hat{f}$, the lower the (squared) bias but the higher the variance.

Bias-Variance Trade-off
Examples with Simulated Data

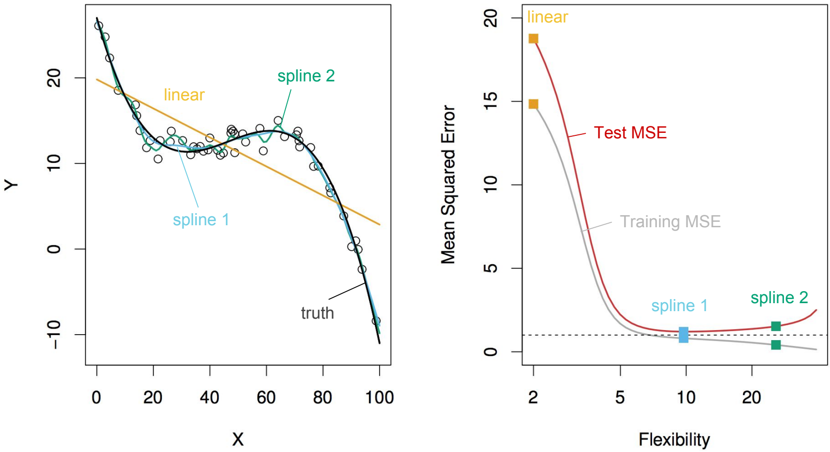
A) 3 model estimates and their MSEs

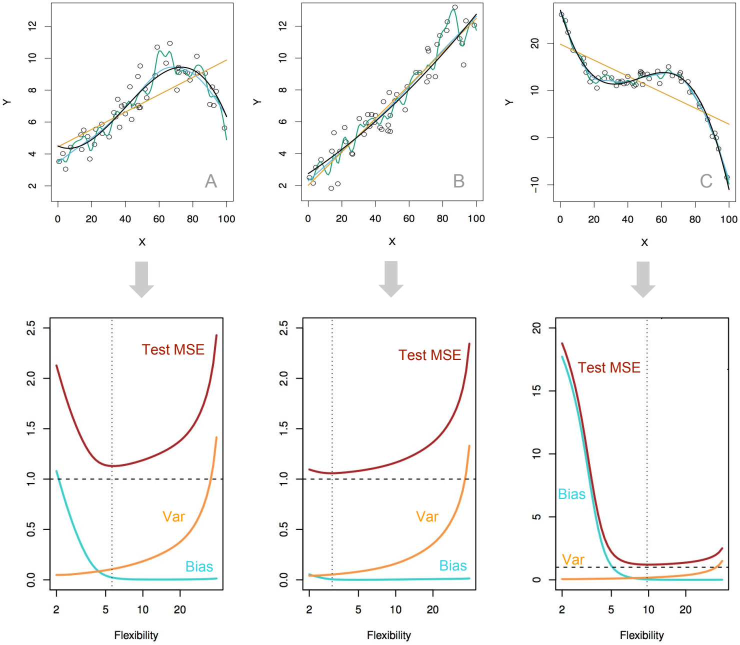


B) 3 model estimates and their MSEs



C) 3 model estimates and their MSEs





Bias-Variance Trade-off

Bibliopgraphy

- **Statistical Regression and Classification** by Norman Matloff (2017). CRC Press.
- **Statistical Learning from a Regression Perspective** by Richard Berk (2008). *Chapter 1: Regression Framework*. Springer.
- **Data Mining: Practical Machine Learning Tools and Techniques** by Ian Witten and Eibe Frank (2005). Elsevier.
- **Modern Multivariate Statistical Techniques** by Julian Izenman (2008). *Chapter 5: Prediction Accuracy and Model Assessment*. Springer.