

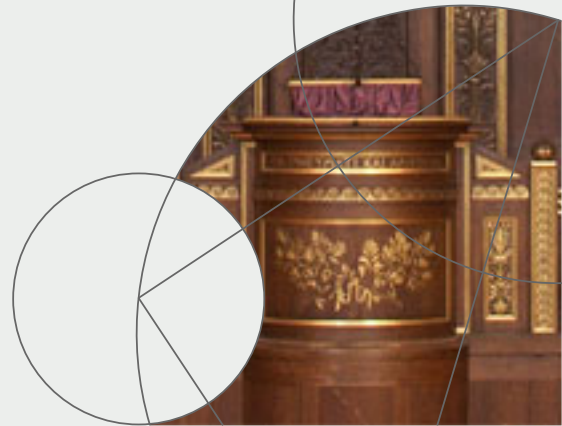


STEVENS INSTITUTE OF TECHNOLOGY



Machine Learning Fundamentals: Bias vs. Variance

Rensheng Wang,
<https://sit.instructure.com/courses/34886>









Bias vs. Variance

- Whenever we discuss model prediction, its important to understand prediction errors (bias and variance).
- There is a tradeoff between a models ability to minimize bias and variance.
- Gaining a proper understanding of these errors would help us not only to build accurate models but also to avoid the mistake of **overfitting** and **underfitting**.
- ☞ **Bias:** Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- ☞ **Variance:** Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.



Bias vs. Variance

-  **Bias:** Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
 -  Model with high bias pays very little attention to the training data and oversimplifies the model.
 -  It always leads to high error on training and test data.
-  **Variance:** Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
 -  Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasnt seen before.
 -  As a result, such models perform very well on training data but has high error rates on test data.



Bias vs. Variance : Mathematically

- Let the variable we are trying to predict as Y and other covariates as X . We assume there is a relationship between the two such that

$$Y = f(X) + e$$

Where e is the error term and its normally distributed with a mean of 0.

- We will make a model $\hat{f}(X)$ of $f(X)$ using linear regression or any other modeling technique. So the expected squared error at a point x is

$$Err(x) = E \left[\left(Y - \hat{f}(x) \right)^2 \right]$$

where $E[\cdot]$ is the expectation.

- As $Y = f(x) + e$, the $Err(x)$ can be further decomposed as

$$Err(x) = E \left[\left(f(x) + e - \hat{f}(x) \right)^2 \right] = E \left[\left(f(x) - \hat{f}(x) \right)^2 + e^2 + 2e[f(x) - \hat{f}(x)] \right]$$



Bias vs. Variance : Mathematically

□ The $Err(x)$ can be further decomposed as

$$\begin{aligned} Err(x) &= E \left[\left(f(x) - \hat{f}(x) \right)^2 + 2e[f(x) - \hat{f}(x)] + e^2 \right] \\ &= E \left[\left(f(x) - \hat{f}(x) \right)^2 + 2e[f(x) - \hat{f}(x)] \right] + \sigma_e^2 \end{aligned}$$

where $E[e^2] = \sigma_e^2$.

□ The noise term e is independent of $f(x)$ and $\hat{f}(x)$, then the expectation $E \left[2e \left(f(x) - \hat{f}(x) \right) \right] = 0$.

$$\begin{aligned} Err(x) &= E \left[\left(f(x) - \hat{f}(x) \right)^2 + 2e[f(x) - \hat{f}(x)] \right] + \sigma_e^2 \\ &= E \left[\left(f(x) - \hat{f}(x) \right)^2 \right] + 0 + \sigma_e^2 \end{aligned}$$



Bias vs. Variance : Mathematically

■ Reform the term $E[(f(x) - \hat{f}(x))^2]$

$$\begin{aligned}
 & E \left[\left(f(x) - \hat{f}(x) \right)^2 \right] \\
 &= E \left[\left(\hat{f}(x) - f(x) \right)^2 \right] \\
 &= E \left[\underbrace{\left(\hat{f}(x) - E[\hat{f}(x)] \right)}_{\text{Variance}} + \underbrace{\left(E[\hat{f}(x)] - f(x) \right)}_{\text{Bias}} \right]^2 \\
 &= E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right)^2 \right] + \left(E[\hat{f}(x)] - f(x) \right)^2 + 2E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right) \left(E[\hat{f}(x)] - f(x) \right) \right] \\
 &= \text{Variance} + \text{Bias}^2 + 2E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right) \left(E[\hat{f}(x)] - f(x) \right) \right] \\
 &\text{where Variance} = E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right)^2 \right] \text{ and Bias} = E[\hat{f}(x)] - f(x).
 \end{aligned}$$

■ Note in the above equation, the bias ($E[\hat{f}(x)] - f(x)$) is a deterministic term and the expectation operation vanished over the bias.



Bias vs. Variance : Mathematically

- We further simplify the term,

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

- As aforementioned, the bias term ($E[\hat{f}(x)] - f(x)$) is independent of the expectation operator, we can move it out of the expectation,

$$\begin{aligned} & E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right) \left(E[\hat{f}(x)] - f(x) \right) \right] \\ &= E \left[\left(\hat{f}(x) - E[\hat{f}(x)] \right) \right] \cdot \left(E[\hat{f}(x)] - f(x) \right) \\ &= \left(E[\hat{f}(x)] - E[\hat{f}(x)] \right) \cdot \left(E[\hat{f}(x)] - f(x) \right) = 0 \end{aligned}$$

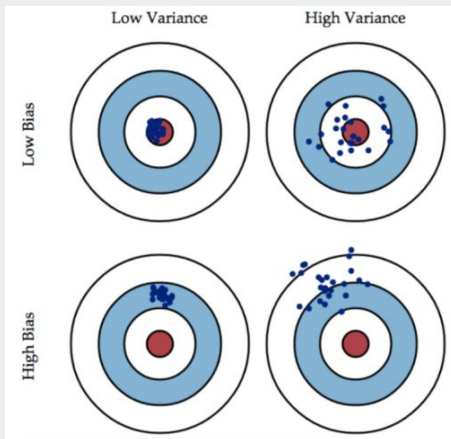
- Overall we have the error term

$$\text{Err}(x) = E \left[\left(f(x) - \hat{f}(x) \right)^2 \right] + \sigma_e^2 = \text{Variance} + \text{Bias}^2 + \sigma_e^2$$



Bias vs. Variance : Tradeoff

- In the error term, the noise variance σ_e^2 is irreducible and therefore we can only work on the bias and variance terms.



Underfitting vs. Overfitting

- ❑ In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data.
- ❑ These models usually have high bias and low variance.
- ❑ It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data.
- ❑ For instance, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.



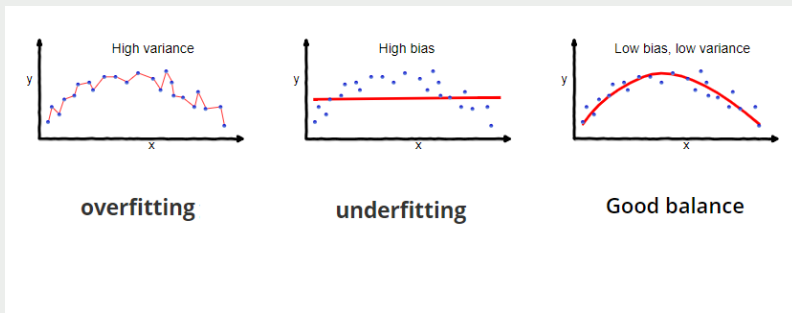
Underfitting vs. Overfitting

- ❑ In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data.
- ❑ It happens when we train our model a lot over noisy dataset.
- ❑ These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.



Underfitting vs. Overfitting

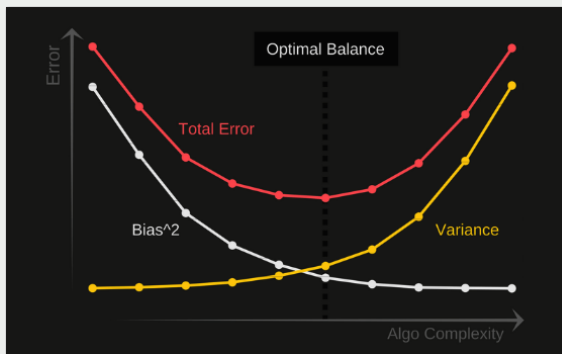
- If our model is too simple and has very few parameters then it may have high bias and low variance.
- On the other hand if our model has large number of parameters then its going to have high variance and low bias.
- So we need to find the right/good balance without overfitting and underfitting the data. This tradeoff in complexity is why there is a tradeoff between bias and variance.



Bias vs. Variance: Tradeoff

- To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$



Underfitting vs. Overfitting

- ❑ Let us assume we have both training and testing data sets.
- ❑ Let us assume in a classification example, the best model, i.e., the benchmark model, whatever using human or machine, can reach 0% error rate.
- ❑ If the model learned on training data has an error rate $\gg 0\%$, for instance, 15%, then there is a big bias, i.e., underfitting.
- ❑ If the learned model applies for the testing data has an error rate 30%, then there is a big variance at this model.
- ❑ Comparing both training and testing data error rates with the benchmark model, this learning method has both big bias and big variance.
- ❑ If the same model applies for the testing data has an error rate 16%, then there is a big bias but small variance at this model.



Underfitting vs. Overfitting

- Let us assume we have both training and testing data sets.
- Let us assume in a classification example, the best model, i.e., the benchmark model, whatever using human or machine, can reach 0% error rate.

Training Error Rate	Testing Error Rate	Bias vs. Variance	Note
15 %	30 %	big bias, big variance	Underfitting
15 %	16 %	big bias, small variance	Underfitting
1 %	15%	small bias, big variance	Overfitting
1%	2%	small bias, small variance	

