Demystifying the Bias-Variance Tradeoff

Ashwin Rao

August 7, 2017

1 Motivation and Overview

The *Bias-Variance Tradeoff* is perhaps the most important concept to learn for any student getting initiated in Machine Learning. Unfortunately, it is not appreciated adequately by many students who get caught up in the mechanics of advanced Machine Learning models/algorithms and don't realize that many of the pitfalls in the models they build are due to either too much Bias or too much Variance. In this note, I will explain this tradeoff by highlighting the probabilistic elements in the derivation of the formula governing the tradeoff, will explain how to interpret the tradeoff, and will finally introduce the concept of *Capacity* that plays a key role in actually "playing the tradeoff".

2 Understanding the Probabilistic Aspects

I think the crux of the *Bias-Variance tradeoff* is lost on many students because the probabilistic aspects of the setting under which the tradeoff operates is not explained properly in most textbooks or by teachers. Let us consider a supervised Machine Learning problem where the set of features are denoted by X and the supervisory variable is denoted by Y. To understand the setting intuitively, let us look at a simple example: Say X consists of the Age, Gender and Country of a person and Y is the Height of the person. Here, we are in the business of predicting the Height from the value of the (Age, Gender, Country) 3-tuple, but we need to understand that for a fixed (Age, Gender, Country) tuple, the Height (as seen in the data) will be spread over a range. Hence, we talk about Height as a probability distribution conditional on the value of the (Age, Gender, Country) tuple. This is really important to understand - that Y given X (denoted $Y \mid X$) is a random variable. Note that since this is conditional on X, we need to treat the conditional probability of $Y \mid X$ as a function of X (meaning the probability distribution of Y depends on X).

In this setting, we denote the Expectation and Variance of the conditional random variable $Y \mid X$ as $\mu(X)$ and $\sigma^2(X)$ respectively. Now let's say we build a model with training data set T whose predicted value (of the supervisory variable) is denoted as $\hat{Y} = \hat{f}_T(X)$. The subscript T is important - it means that the model prediction function $\hat{f}$

depends on the training data set T. The intuition here is that we aim to get $\hat{f}_T(X)$ reasonably close to $\mu(X)$, i.e., we hope that our model's prediction for a given X will be close to the conditional expectation of Y given X.

The Bias-Variance tradeoff is a statement about expectations under two different (and independent) sources of randomness:

- The randomness associated with Y conditioned on X. We will refer to this source of randomness by subscripting with the notation $Y \mid X$.
- The randomness associated with the choice of training data set T which in turn results in randomness in the model-prediction function $\hat{f}_T$. We will refer to this source of randomness by subscripting with the notation T

Note that that these two sources of randomness $Y \mid X$ and T are independent

3 Expected Prediction Error for a Test Data Point

The Expected Prediction Error (EPE) of the model on a test data point (x, y) is defined as:

$$\begin{aligned}
EPE_{(Y|X),T}(x) &= E_{(Y|X),T}[(\hat{f}_T(x) - y)^2] \\
&= E_{(Y|X),T}[\hat{f}_T^2(x) + y^2 - 2 \cdot \hat{f}_T(x) \cdot y] \\
&= E_T[\hat{f}_T^2(x)] + E_{Y|X}[y^2] - 2E_{(Y|X),T}[\hat{f}_T(x) \cdot y]
\end{aligned}$$

Note that:

$$E_{Y|X}[y^2] = \mu^2(x) + \sigma^2(x)$$

$$E_{(Y|X),T}[\hat{f}_T(x) \cdot y] = E_T[\hat{f}_T(x)] \cdot E_{Y|X}[y] = E_T[\hat{f}_T(x)] \cdot \mu(x)$$

(because of independence of the two sources of randomness)

Hence,

$$EPE_{(Y|X),T}(x) = E_T[\hat{f}_T^2(x)] + \mu^2(x) + \sigma^2(x) - 2 \cdot E_T[\hat{f}_T(x)] \cdot \mu(x)$$

4 Bias and Variance

Before we state the definitions of Bias and Variance in precise equational language, let us understand them intuitively. Both Bias and Variance refer to the probabilistic nature of the model's forecast for a given test data point x, with the probabilities governed by the random choices in selecting the training data set T.

Bias of a model refers to the "gap" between:

- The model's expected prediction of the supervisory variable (corresponding to the given test data point x). Note that this expectation is over probabilistic choices of training data set T.
- The expected value of the supervisory variable (corresponding to the given test data point x) that actually manifests in the data. Note that this expectation is over the probability distribution of Y given X (notationally, $Y \mid X$).

Variance of a model refers to the "expected squared deviation" of the model's prediction of the supervisory variable (corresponding to the given test data point x) around the expected prediction of the supervisory variable. Note that this "expected squared deviation" is over probabilistic choices of training data set T and is meant to measure the degree of "fluctuation" (or you may want to call it "instability") in the model's prediction (due to variations in the choice of the training data set T).

Now we precisely define the Bias and Variance of the model $\hat{f}_T$ when it makes a prediction for the test data point x.

$$\begin{aligned} Bias_T(x) &= E_T[\hat{f}_T(x)] - \mu(x) \\ Variance_T(x) &= E_T[(\hat{f}_T(x) - E_T[\hat{f}_T(x)])^2] = E_T[\hat{f}_T^2(x)] - (E_T[\hat{f}_T(x)])^2 \end{aligned}$$

$$\begin{aligned} & Bias_T^2(x) + Variance_T(x) \\ &= (E_T[\hat{f}_T(x)])^2 + \mu^2(x) - 2 \cdot E_T[\hat{f}_T(x)] \cdot \mu(x) + E_T[\hat{f}_T^2(x)] - (E_T[\hat{f}_T(x)])^2 \\ &= \mu^2(x) - 2 \cdot E_T[\hat{f}_T(x)] \cdot \mu(x) + E_T[\hat{f}_T^2(x)] \\ &= EPE_{(Y|X),T}(x) - \sigma^2(x) \end{aligned}$$

In other words,

$$EPE_{(Y|X),T}(x) = Bias_T^2(x) + Variance_T(x) + \sigma^2(x)$$

5 Interpreting the Tradeoff

So we can see that the Expected Prediction Error of a model (built from random training data) on a test data point (x, y) (that is governed by the conditional randomness $Y \mid X$) is composed of 3 parts:

- $\sigma^2(x)$: Remember that models are in the business of predicting $E[y \mid x] = \mu(x)$ whereas the EPE compares the model prediction to $y \mid x$ (not to $E[y \mid x]$), so the conditional (on x) variance around $E[y \mid x]$ (i.e. $\sigma^2(x)$) will always exist in the EPE and is not reducible by any model. So we will focus the rest of the discussion in this note on how to interpret the remaining two terms, and finally how to control them in "playing the tradeoff".

- $Bias_T^2(x)$: This term has to do with the fact that the model $\hat{f}_T$ might not adequately capture the complexity of the function defined by the conditional expectation of Y given X. Under-parameterized models are too simple to capture the richer structure of data and suffer from high $Bias_T$. An example would be a simple linear regression trying to capture structure of data that has say an exponential relationship between X and Y. On the other hand, a model that captures the essential complexity in the relationship between X and Y will have low/no bias.

- $Variance_T(x)$: This term has to do with the fact that the randomness of (variability in) the choice of the training data set T results in variability in the predictions of the model $\hat{f}_T$. Over-parameterized models are essentially unstable models and suffer from high $Variance_T$. An example would be a nearest neighbors model that has too many degrees of freedom (too many parameters) trying to fit perfectly to the training data. On the other hand, a model that doesn't fit the training data too tightly will have low variance.

6 Capacity

The key in building a good machine learning model is to find the right *effective level of parameterization* that balances the two effects of Model Bias and Model Variance (trading one against the other). Thankfully, we have a precise technical concept called *Capacity* of a model that intuitively refers to the *effective level of parameterization*. You can also think of *Capacity* as the "span" of functions that can be captured by the model (hence, the term "Capacity"), or you can simply think of it as the "Model Complexity". A linear regression model's "span" would be the set of all linear functions. A polynomial model's "span" would be the set of all polynomials (up to its specified degree). A nearest neighbor model would have as much capacity as the set of training data points (since there is a parameter for each data point), and hence it would typically have high capacity (assuming we have plenty of training data points). Regularization is a technique which tones down the *Capacity*.

But how does *Capacity* serve as a mechanism to actually play the *Bias-Variance Tradeoff*? Let's say you start with a simple linear regression model. We know this has low *Capacity*. Training data error would be large if the data itself is fairly non-linear. Now let's keep increasing the *Capacity*. We will find that training data error will keep reducing because the richer structure of permitted functions will aim to fit the training data more and more precisely. But there is no free lunch - the test data error (EPE) will increase if you increase the *Capacity* too much. There will be an *Optimal Capacity* somewhere in between where the EPE is the lowest. This is where you have found the right balance between the Model Bias and Model Variance. If the *Capacity* is lower than this *Optimal Capacity*, you run the risk of underfitting - high Model Bias and low Model Variance. If the *Capacity* is higher than this *Optimal Capacity*, you run the risk of overfitting - low Model

Bias and high Model Variance. So, *Capacity* is a mechanism to play Model Bias against Model Variance in an attempt to find the right balance.

The other point to note is that *Capacity* has a connection with the number of training data points. If you have a larger number of training data points, the *Optimal Capacity* will be tend to be larger (until the *Optimal Capacity* plateaus - as it eventually achieves sufficient complexity to solve the problem).

This note did not go into the mathematical specification of the technical term *Capacity* as I wanted to keep this to an introductory content, but mathematically advanced readers are encouraged to understand the technical term *Capacity* by looking up the definition of *Vapnik-Chervonenkis dimension* and how it is used to bound the gap between the test data error and the training data error. Sadly, the VC dimension is not very useful in practical Machine Learning models, but it serves as a great metric to understand and appreciate the significance of *Capacity* and its connection with the *Bias-Variance Tradeoff*.