Bias and Variance in Machine Learning

Jesse Hoey
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, CANADA, N2L3G1
jhoey@cs.uwaterloo.ca

Bias and variance are key concepts in machine learning. This note gives an intuitive account of these concepts. We start by defining three key dimensions of the space of learning algorithms.

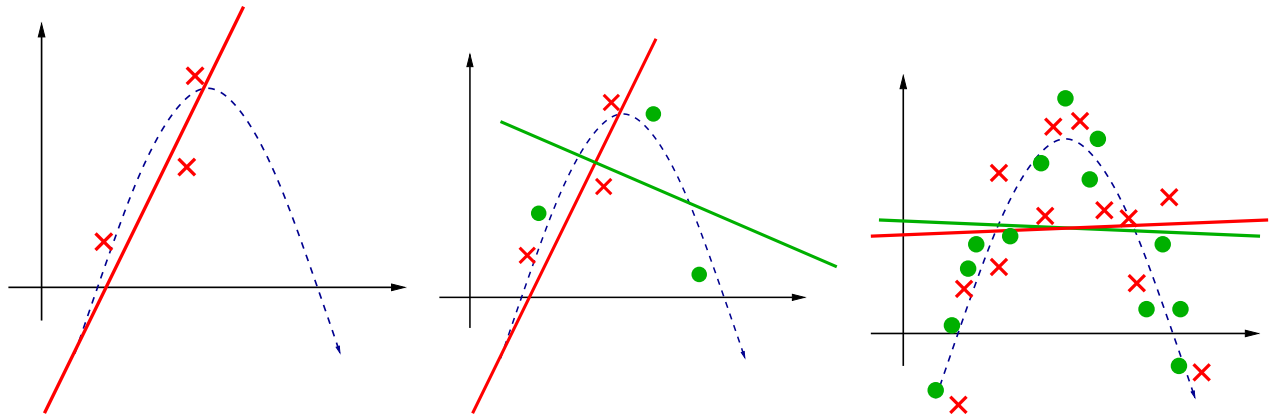
- **Bias**: describes the (inverse of the) complexity of the model space that is being learned. High bias models have few parameters, and so are simple (not complex) and have low *capacity*. That is, the set of datasets to which they are applicable is small. For example, a straight line (with two parameters) can only fit data that lies in a straight line. Low bias models have many parameters, and so are complex and have high capacity. These low bias models can be fit to many more datasets. In some sense, the lowest bias model is the dataset itself (a template-based model), and the "P" and "N" agents are examples of such models (see Box 1).

In a binary classification task, the "P" agent claims the negative examples seen are the only negative examples. Every other instance is positive, while the "N" learner claims the positive examples seen are the only positive examples. Every other instance is negative. Both agents correctly classify every training example, but disagree on every other example. They are massively overfit to the training data and have very low bias (fits every training set perfectly) and very high variance (opposing classification of every single test instance).

Box 1

- **Data**: how much data you have. This is loosely defined, and we will see that what is important is how much data you have *relative* to the complexity of the model that is being learned.
- **Variance**: describes how much difference there is between learned models on different datasets of a fixed size. That is, if you gather a set of data and learn a model (call this model "a") and then do it again (gather a second set of data of the same size as the first and learn a new model, call it model "b"), how different will model "b" be from model "a"? This is the variance.

Consider the function shown with a dotted line in Figure 2(a). Suppose we draw a dataset of 3 data points from this function, shown with red $\times$ in Figure 2(b), and fit a straight line to it. This is a very high bias model (a simple function) and it fits these three datapoints reasonably well. However, it is not a good model of the underlying data generating process! To see why, let's draw a second dataset with another 3 points (shown with green circles in Figrue 2(b)). Fitting the same high bias model yields a *very* different solution. This big difference is a measure of



Box 2: (a) data generating process (unknown); (b) with little data, fitting high bias models (straight lines) gives high variance solutions; (c) with more data, high bias models give lower variance.

the *variance* in this learning procedure, which we see in this case is very high. To mitigate the high variance, we can sample more data, shown in Figure 2(c). Now, we see the same high bias models yield roughly the same result for two different (random) samples from the data generating function. So we see that increasing the size of the dataset while keeping the bias constant yields a lower variance fit. This is because the bias of the model relative to the dataset has in fact gone up, so the variance can decrease.

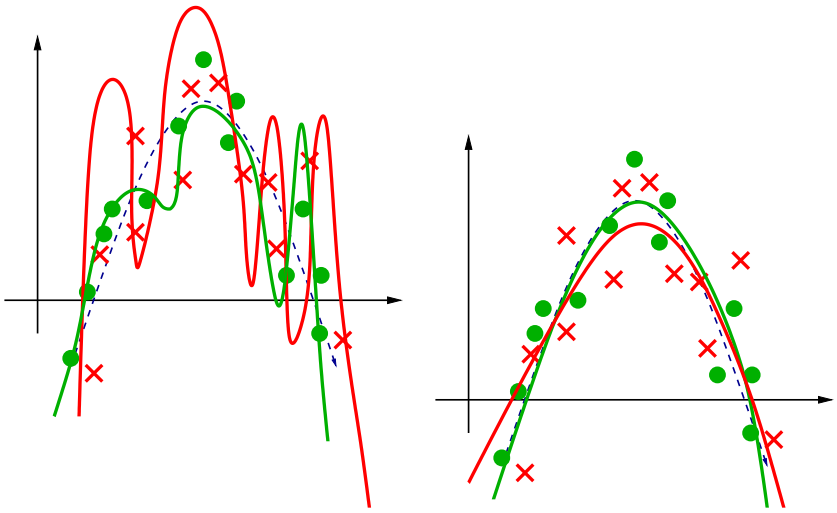
Now let us try increasing the complexity of the model (decreasing its bias). As shown in Figure 3(a), the lower bias model may fit the data much better (come closer to all the training data than in Figure 2(c)), but yields a high variance solution. Finding the right balance between bias and variance can yield satisfactory results with low variance (Figure 3(b)).

Overall, to have low variance, high bias models (with respect to the dataset size) are needed. In fact, a formal link between model complexity and amount of data is the Vapnik–Chervonenkis (VC) dimension, defined as the cardinality of the largest set of points that the learning function can "shatter" [VC15]. In simple terms, the more data you have, the more complex the function needs to be to fit it well. As the bias is decreased (model made more complex with respect to the size of the dataset, or dataset size decreased), in general the variance will increase. Conversely, if we attempt to decrease variance by sampling more data (or making the model less complex), then the bias will increase relative to the dataset.

In machine learning, one ultimately is looking for a low bias and low variance model. This is one that makes little assumption about the form of the underlying data generating process, and consistently yields the same result regardless of the dataset gathered. Such a model, as it is robust to the training data, will *generalize* well. However, the fundamental tradeoff between bias and variance shows that, if trying to decrease bias, variance increases, and if trying to decrease variance, bias increases.

There is, however, a slight wrinkle in this story. Recent results have shown that in the *overparameterized* regime,[1] this effect is reversed, and lowering bias starts to *decrease* variance [NMB+18]. This is because, in fact, variance arises for two reasons. First, variance due to sampling is what we have described above: if different datasets are gathered, then the results are differ-

[1] Overparametrized models are more complex than required by the VC dimension. Thus, a polynomial of degree two is overparameterized for data that lie in a straight line.



(a) (b)

Box 3: (a) low bias models overfit, variance is high; (b) just the right bias and variance

ent. This variance is reduced by gathering more data (or equivalently increasing bias relative to the dataset size). This type of variance continues to *increase* as bias is decreased beyond the overparametrization boundary defined by the VC dimension. The second type of variance is due to optimization, and arises because the process of fitting a model to a set of data is not deterministic for complex models, and may yield different results each time. This type of variance apparently starts to *decrease* in the overparameterized regime. Intuitively, we can see why because when the model is overparameterized, it can find more parameter settings that fit a single dataset with the same precision, which means it will be easier for the optimization to find the same solution as before. The sum of both variances can be shown to decrease (so the variance due to optimization plays a bigger role than the variance due to sampling once the model is overparameterized). These results are still being investigated.

References

[NMB+18] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *CoRR*, abs/1810.08591, 2018.

[VC15] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In Vladimir Vovk, Harris Papadopoulos, and Alexander Gammerman, editors, *Measures of Complexity: Festschrift for Alexey Chervonenkis*, pages 11–30. Springer International Publishing, Cham, 2015.