LECTURE 1 : Paper Work

BIAS (in neural networks) **Komal Mehra (15mi413)**

How is it possible to learn in any meaningful sense?

The answer is that learning is possible in the restricted sense that the fitted model will probably perform approximately as well on new data as it did on the training data. Once an appropriate error function E is chosen for the problem under consideration (e.g. sum of squared errors in linear regression), we can define two distinct performance measures of interest. The in-sample error, Ein, and the out-of-sample or generalization error, Eout. Both metrics are required due to the distinction between fitting and predicting. This raises a natural question: Can we say something general about the relationship between Ein and Eout? Surprisingly, the answer is 'Yes' . We can in fact say quite a bit. This is the domain of statistical learning theory.

The first schematic, shows the typical out-of-sample error, Eout, and in-sample error, Ein, as a function of the amount of training data. In making this graph, we have assumed that the true data is drawn from a sufficiently complicated distribution, so that we cannot exactly learn the function f(x). Hence, after a quick initial drop (not shown in figure), the in-sample error will increase with the number of data points, because our models are not powerful enough to learn the true function we are seeking to approximate. In contrast, the out-of-sample error will decrease with the number of data points. As the number of data points gets large, the sampling noise decreases and the training dataset becomes more representative of the true distribution from which the data is drawn. *For this reason, in the infinite data limit, the in-sample and out-of-sample errors must approach the same value, which is called the "bias" of our model. The bias represents the best our model could do if we had an infinite amount of training data to beat down sampling noise.* The bias is a property of the kind of functions, or model class, we are using to approximate f(x). In general, the more complex the model class we use, the smaller the bias. However, we do not generally have an infinite amount of data. For this reason, to get best predictive power it is better to minimize the out-of-sample error, Eout, rather than the bias. As shown in Fig. 1, Eout can be naturally decomposed into a bias, which measures how well we can hypothetically do in the infinite data limit, and a variance, which measures the typical errors introduced in training our model due to sampling noise from having a finite training set.

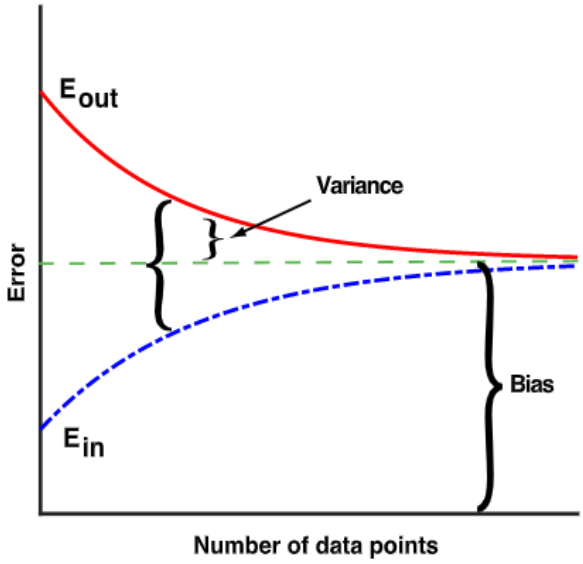

Fig. 1. Schematic of typical in-sample and out-of-sample error as a function of training set size. The typical in-sample or training error,Ein, out-of-sample or generalization error,Eout, bias, variance, and difference of errors as a function of the number of training data points.

The second schematic, shown in Fig.2, shows the out-of sample, or test, error Eout as a function of ''model complexity''. Model complexity is a very subtle idea and defining it precisely is one of the great achievements of statistical learning theory. In many cases, model complexity is related to the number of parameters we are using to approximate the true function f(x).1In the example of polynomial

regression discussed above, higher-order polynomials are more complex than the linear model. If we consider a training data set of a fixed size, Eout will be a non-monotonic function of the model complexity, and is generally minimized for models with intermediate complexity. The underlying reason for this is that, even though using a more complicated model always reduces the bias, at some point the model becomes too complex for the amount of training data and the generalization error becomes large due to high variance. Thus, to minimize Eout and maximize our predictive power, it may be more suitable to use a more biased model with small variance than a less-biased model with large variance. This important concept is commonly called the bias–variance tradeoff and gets at the heart of why machine learning is difficult.

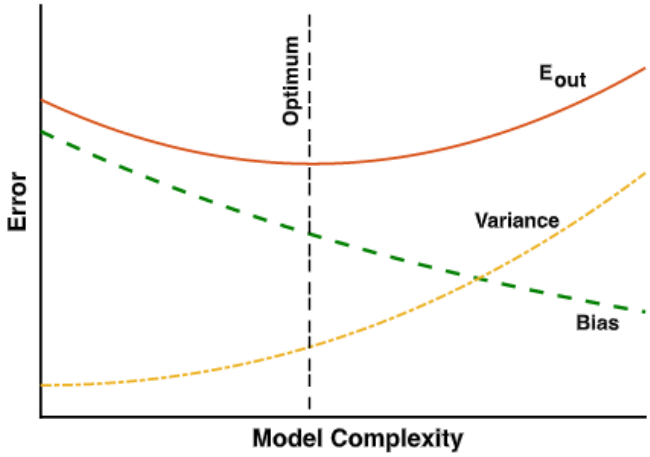

Fig. 2. Bias–Variance tradeoff and model complexity

Another way to visualize the bias–variance tradeoff is shown in Fig. 3. In this figure, we imagine training a complex model (shown in green) and a simpler model (shown in black) many times on different training sets of a fixed size N. Due to the sampling noise from having finite size datasets, the learned models will differ for each choice of training sets. In general, more complex models need a larger amount of training data. For this reason, the fluctuations in the learned models (variance) will be much larger for the more complex model than the simpler model. However, if we consider the asymptotic performance as we increase the size of the training set (the bias), it is clear that the complex model will eventually perform better than the simpler model. Thus, depending on the amount of training data, it may be more favourable to use a less complex, high-bias model to make predictions.

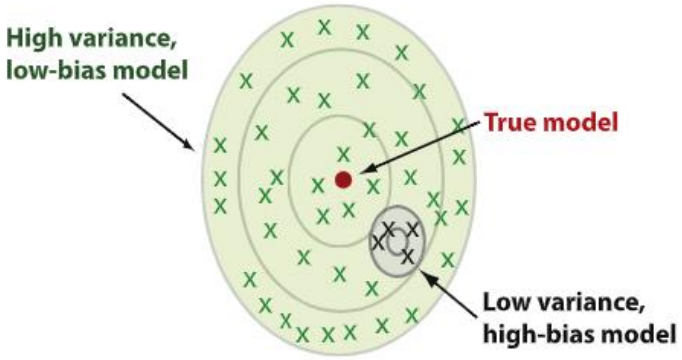

Fig. 3. Bias–Variance tradeoff

REFERENCE: B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Julien, I. Mitliagkas. “A modern take on the bias-variance tradeoff in neural networks”, October 2018