The Bias-Variance Dilemma

Raúl Rojas

February 10, 2015

Abstract

This tutorial explains the so-called *bias-variance dilemma*, also called the *bias-variance tradeoff*, which arises when fitting a function to experimental data. Complex models have a tendency to overfit the data, which is noisy in general (the models will then exhibit high variance or variability). However, simplistic models could lack the flexibility needed to approximate complex processes (they then have high bias). A compromise has to be found between bias and variance, hence the dilemma

1 Motivation

We review the so-called *bias-variance dilemma* using two examples shown in Fig. 1. In the lower diagram we assume that we have been provided with a set of experimental data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ (blue points in the figure). The data comes from a nonlinear process represented by the function $f(x)$. However, we are ignorant of this function underlying the process and the given experimental data contains noise. That is, every y_i is of the form $f(x_i) + \sigma$, where σ represents white noise. We decide to fit linear functions to the data.

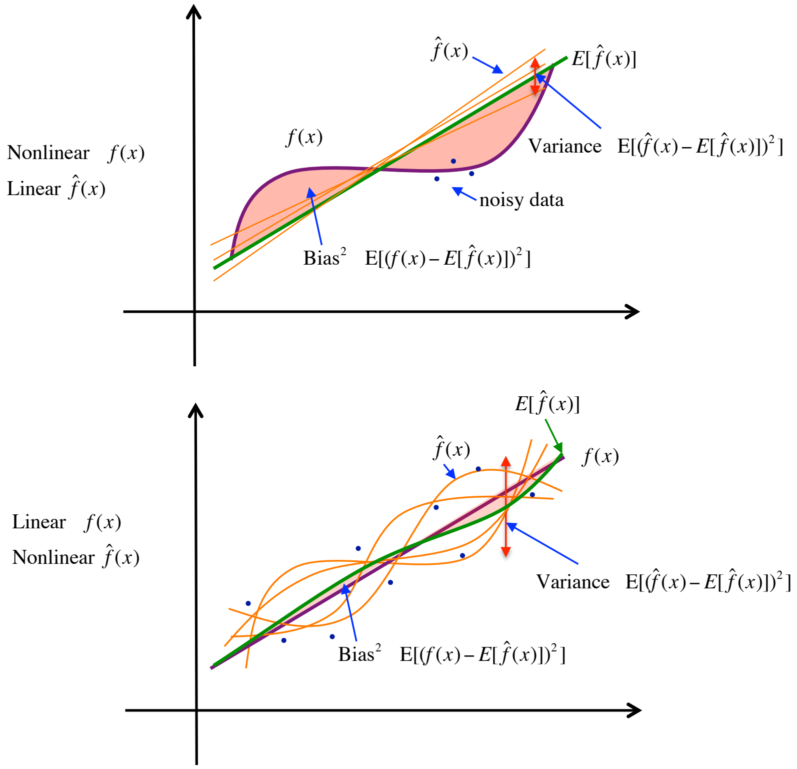
$\hat{f}(x)$
$E[\hat{f}(x)]$
$f(x)$
Nonlinear $f(x)$
Linear $\hat{f}(x)$
Variance $\mathrm{E}[(\hat{f}(x)-E[\hat{f}(x)])^2]$
noisy data
Bias^2 $\mathrm{E}[(f(x)-E[\hat{f}(x)])^2]$
$E[\hat{f}(x)]$
$\hat{f}(x)$
$f(x)$
Linear $f(x)$
Nonlinear $\hat{f}(x)$
Variance $\mathrm{E}[(\hat{f}(x)-E[\hat{f}(x)])^2]$
Bias^2 $\mathrm{E}[(f(x)-E[\hat{f}(x)])^2]$

Figure 1: Two examples for the bias-variance dilemma.

Assume that we are provided with a collection of data sets from the same process (that is, we repeat the fitting experiment with different representative sets of data coming from the same experiment). Our linear fit $\hat{f}(x)$ will be a line for each data set, and we will obtain several of them, one for each set (orange lines in Fig. 1, top diagram). If we compute the average of all those lines, that is, the expected value of the fitting function, we obtain $E[\hat{f}(x)]$ (the green line in the diagram). Since the underlying process is nonlinear, in general there will be a difference between the function $f(x)$ and the average fit $E[\hat{f}(x)]$ (red surface in the figure). The sum of squares of all those differences is what is called the "squared bias" of the fitting function: That is, since the fitting function is just a line, it is "biased" toward linear functions. Numerically it is computed as the expected squared value of the difference between $f(x)$ and the average fit $E[\hat{f}(x)]$, that is $E[(f(x) - E[\hat{f}(x)])^2]$.

Since we compute alternative fits for the different data sets, there is a difference between each individual fit $\hat{f}(x)$ and the average of all fits $E[\hat{f}(x)]$. We call the expected value of the squared difference the "variance" of the model, given by $E[(\hat{f}(x) - E[\hat{f}(x)])^2]$. As we can see from the top diagram in Fig. 1, the simplicity of our linear model leads to a high bias for the nonlinear process $f(x)$, but a somewhat small variance, since all the computed linear functions do not differ much for the different data sets.

One additional source of error for the fit is the noise in the y-values. We call σ^2 the expected value of the squared difference between y and $f(x)$.

The lower diagram in Fig. 1, shows the opposite example. Now the underlying process $f(x)$ is linear (purple line). The functions we are fitting are nonlinear and have several degrees of freedom (they could be polynomials of degree 5, for example). Since the data is noisy, each single fit overfits the data. Several fits to alternative data sets will look very different. Therefore, their "variance" is high. But their average will come near to the line representing the underlying process (since polynomials of degree 5 contain lines as a special case). Given enough data, the fitted functions $\hat{f}(x)$ will generate a smoothed out $E[\hat{f}(x)]$. Now the "bias" between the process $f(x)$ and the model $E[\hat{f}(x)]$ is low.

2 The algebra

The algebra for obtaining the relationship between all these three things (bias, variance, and noise) is straightforward. We are interested in the expected quadratic error of a single fit

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

Since the noise is uncorrelated to the goodness of the fit, when we expand the quadratic function and compute the sum of expectations, we will have that $E[2\sigma(f(x)-\hat{f}(x))]$ is equal to zero. Therefore

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

Now, we can rewrite $E[(y-\hat{f}(x))^2]$ as

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

and using the same argument, that the differences $f(x)-E[\hat{f}(x)]$ and $E[\hat{f}(x)]-\hat{f}(x)$ are uncorrelated, we can ignore the cross terms after expanding the square and thus obtain

$$\underbrace{E[(y-\hat{f}(x))^2]}_{\text{Expected sq error}} = \underbrace{E[(f(x)-E[\hat{f}(x)])^2]}_{\text{Bias}^2} + \underbrace{E[((\hat{f}(x)-E[\hat{f}(x)])^2]}_{\text{Variance of the model}} + \underbrace{\sigma^2}_{\text{Noise}}$$

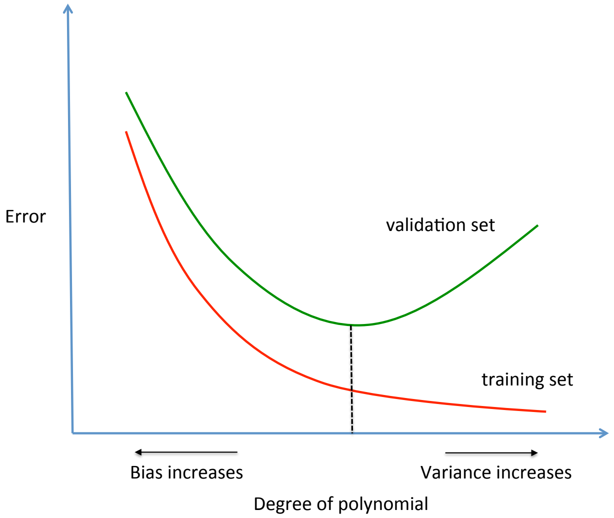

Figure 2: The training error always decreases, as the flexibility of the model is increased. The test error decreases first, but can increase when the complexity of the model is such that it starts to overfit the noisy data.

The interpretation is straightforward, as shown in Fig. 1. In general, we want to have as simple a model as possible (risking a high bias if the process is more complex than the model) but as flexible as needed (risking a high variance if the model has many degrees of liberty). The usual way of balancing both requirements is by dividing the data set in a training and a validation (or test) set. If we are fitting polynomials, for example, we can vary the degree of the polynomial from 1 to 7, fit the approximations using the training set, and test the resulting error with the validation set. We should obtain something like Fig. 2. As the degree of the polynomial advances the error over the training data will always decrease, but the test error can start to increase as soon as the model starts overfitting the data (and the variance of the models obliterates their bias). We also say that simpler models *generalize* better (that is, they have low variance) but can sometimes *underfit* the data, due to their high bias. Finding the right model is all about the compromise between both error contributions, bias and variance.

References

[1] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.