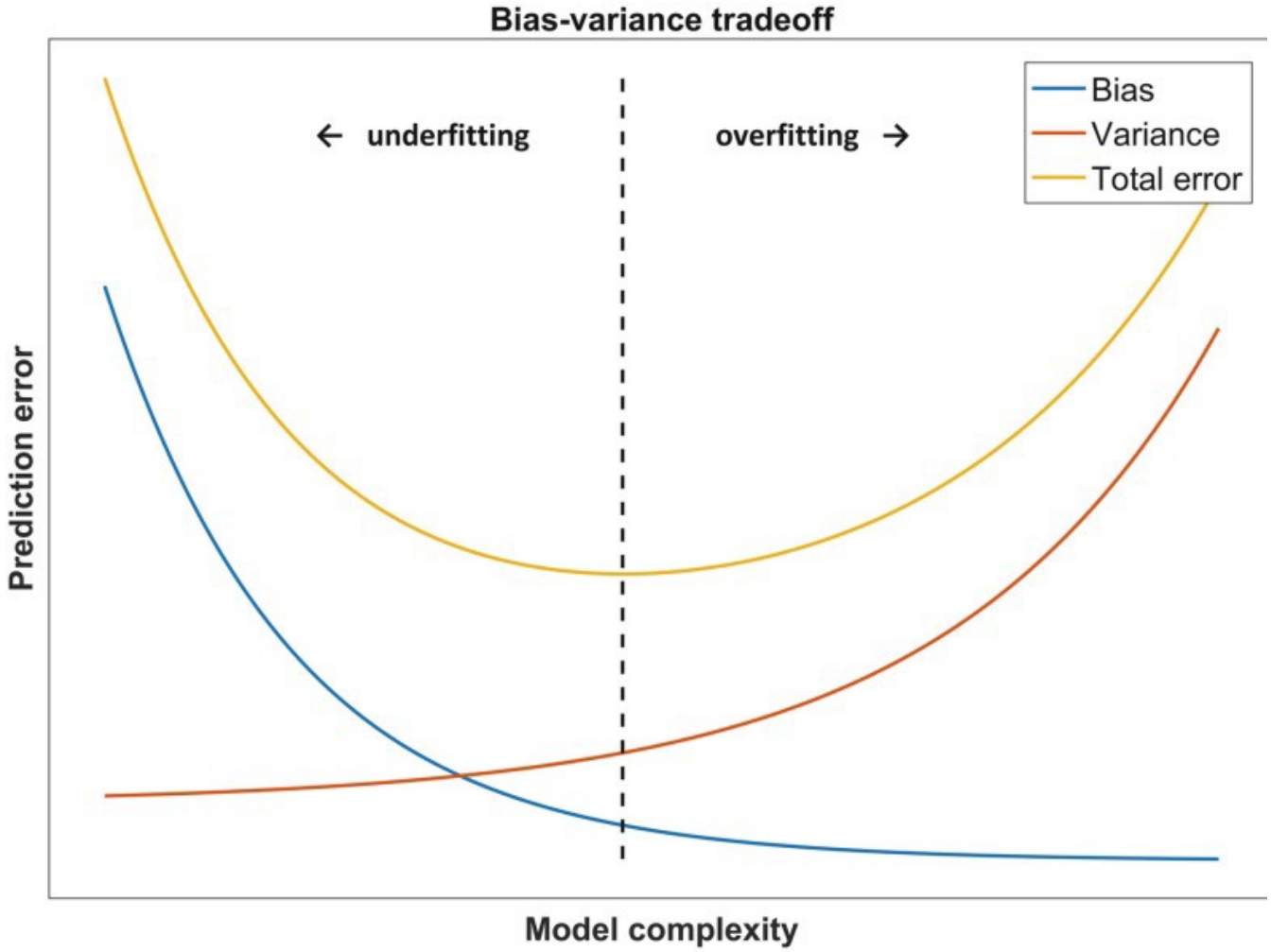
Thoughts and Theory

Simple mathematical derivation of bias-variance error



A simple intuitive figure that represent the prediction (test data error) error as a function of the model complexity. The bias error decreases with model complexity while the variance error increases (Image credit from the book "Fundamentals of Clinical Data Science" under the terms of Creative Commons Attribution 4.0 International License.)

1. Introduction

One of the most important concepts in statistical modelling, data science, and machine learning is that of ***bias-variance error***. This concept is very important because it helps us understand the different

errors that appear in our mathematical modelling when we try to fit the data to predict and make an inference. However, one problem related to this concept is that usually, it is not much clear to the student/researcher working in data science/machine learning, how the bias-variance error relation is derived.

The main reason is related to the fact that many times the bias-variance error (BV error) concept is *taught very superficially* in most learning materials and courses. Very often the term ***bias error*** is introduced as the error that arises in our statistical modelling due to the difference between our selection of the fitting model function to the true model function. This means that if we select a simple fitting function for our statistical modelling when the true function is more complicated, then we are introducing a bias in our selection of the fitting function. On the other hand, the ***variance error*** is introduced as that error in estimating the fitting function to different sample datasets used in our modelling. This means that if we use one particular dataset to fit our selected model function, then if we use a different dataset, our new fitted function for the new dataset might change substantially to that previously found, depending on the sample dataset and its *size*.

The two simple definitions of bias and variance errors given above are accurate to some extent but do not give to the student/researcher an explanation of how these errors arise in statistical modelling. In this article, I derive the BV error relation by using the statistical theory that hopefully will help you better understand the BV error. Here I will assume that the reader knows mathematical analysis and statistical theory. In fact, the purpose of this article is to give a rigorous derivation while trying to keep the mathematical notation as simple as possible.

2. Notations and definitions

Let me start first by introducing some notations that will be useful in what follows. Here, $\boldsymbol{X}$ is the *dependent variable* or *predictor* or *feature* matrix and $\boldsymbol{y}$ is the *independent* or *output variable* vector. Other important notations are the dataset, $D=(\boldsymbol{X}, \boldsymbol{y})$, and the model function $f(\boldsymbol{X}; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the parameter vector of our selected model. For example, in the simple linear regression where we try to fit a linear function, the model parameters can be the intercept and slope of the least square line. The last notation that will use below is the *loss function* or *cost function* $C(\boldsymbol{y}, f(\boldsymbol{X}; \boldsymbol{\theta}))$ which is a measure of model performance on the observations $\boldsymbol{y}$**.** The whole goal of statistical/machine learning is *simply* the following: given the dataset $D=(\boldsymbol{X}, \boldsymbol{y})$, find the parameter vector $\boldsymbol{\theta}$ that minimises the cost function $C(\boldsymbol{y}, f(\boldsymbol{X}; \boldsymbol{\theta}))$. This is the most compact and simple definition of statistical/machine learning. In practice, the dataset is dived into *training* and *test* data for model performance, but I will not go into details because I assume that the reader is familiar with these concepts.

Another important concept that I will use later quite extensively is that of the *mathematical expectation* or *expected value* or simply *expectation* of a generic random variable X. Here X is a generic random variable to not be confused with the above feature matrix $\boldsymbol{X}$**.** For simplicity, here I consider the case of when the random variable X has quantitative values. So, let X have N discrete values

$$X = (x_1, x_2, ..., x_N)$$

Then expectation value of the random variable X is defined as

$$E(X) = x_1 P(X = x_1) + x_2 P(X = x_2) + \ldots + x_N P(X = x_N) = \sum_{i=1}^{N} x_i P(X = x_i), \quad (1$$

where *P(X=x)* is the *discrete probability distribution function* of the random variable *X*. The expectation value *E* has different properties such as*: E(X+Y) = E(X) + E(Y), E(aX) = aE(X), E(XY) = E(X)E(Y)* if *X* and *Y* are random independent variables and *a* a generic real number. Another important quantity is the variance of a random variable *X* which is defined as: Var(*X*)=*E([X - E(X)]²)*, where usually *E(X)* is called the *mean* of *X*. In case the random variable *X* is continuous, then one needs to replace the sum in equation (1) with an integral and *P(X)* with a continuous probability distribution function. The BV relation to be derived below is valid for both discrete and continuous quantitative variables.

---

3. Model decomposition and the cost function

Assume now that we have a given dataset where the true function, $y(\boldsymbol{x})$, that generates the data is given by

$$y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$$
$$E(\epsilon_i) = 0$$
$$E(\epsilon_i \epsilon_j) = \delta_{ij} \sigma^2_{\epsilon_i} \quad (2)$$

where ϵ is the random error or random noise that contributes to the true function $y(\boldsymbol{x})$. *The error ϵ is assumed to be normally distributed with mean zero and standard deviation σ, as shown in equation (2). The different components of the error variable ϵ are also assumed to be*

uncorrelated.

Suppose that we have a given dataset, $D=(\boldsymbol{X}, \boldsymbol{y})$, and we want to perform a typical regression for discrete quantitative input and output variables. In principle, the dataset D can be any type of datasets such as that of the training data or the test data but here I am interested in the *test dataset* because that is the main analysis goal of statistical/machine learning. As is common in statistical/machine learning, it is necessary to evaluate the cost function $C(\boldsymbol{y}, f(\boldsymbol{X}; \boldsymbol{\theta}))$ for a given dataset.

The cost function depends on the type of distance measurement method used and here I will use the typical *Euclidean distance measure(Euclidean metric),* where the cost function can be written as:

$$C = \sum_{i=1}^{N} [y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta})]^2 \qquad (3)$$

The main goal of statistical/machine learning is: given a fixed dataset, find the parameter vector $\boldsymbol{\theta}$ that minimises the cost function C or equivalently*:*

$$\bar{\boldsymbol{\theta}}_D = \arg\min_{\boldsymbol{\theta}} [C(\boldsymbol{y}, f(\boldsymbol{X}; \boldsymbol{\theta}))]$$

If for example a different dataset is used, the cost function $C(\boldsymbol{y}, f(\boldsymbol{X}; \boldsymbol{\theta}))$ would be different, and also the parameter vector $\boldsymbol{\theta}$ that minimises the cost function would be different***. Thus, an important thing to keep in mind is that the cost function and the parameter vector values depend on the dataset.***

Suppose that now we already "learned" the parameter vector $\boldsymbol{\theta}$ from the

training dataset and want to calculate the cost function for the *test dataset***.** In this case, the cost function of the *test dataset*, by using the general definition in equation (3), is given by

$$C_{\text{test}} = \sum_{i=1}^{N} [y_i - f(\boldsymbol{x}_i; \bar{\boldsymbol{\theta}}_D]^2 \qquad (4)$$

One can observe that in equation (4) the cost function of the *test dataset* explicitly depends on the previously learned parameter vector $\boldsymbol{\theta}$ with subscript D.

If you have arrived so far by paying attention to all definitions and to the equation (4), then I must congratulate you for your patience and will. However, there is more to be added since I have not yet derived the BV error expression, so, be patient and keep following.

---

4. Bias-variance error derivation

To derive the BV error, I have to note that it depends on the particular test dataset and on the random error ϵ instance. Keeping this information in mind, now I calculate the *expectation value E(C)* of the *test cost function* in (4) for *different possible test datasets that might be sampled from a population and different error instances*. In this case, the cost function in (4) is a random variable because it implicitly depends on the error ϵ (because of the decomposition in (2)) which is a random variable itself. By using equation (4) in equation (2), I get:

$$E_{D,\epsilon}(C_{\text{test}}) = E_{D,\epsilon}\left(\sum_{i=1}^{N}\left[y_i - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)$$

$$= E_{D,\epsilon}\left(\sum_{i=1}^{N}\left[\underbrace{y_i - f(\boldsymbol{x}_i)}_{\epsilon_i} + f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right) \qquad \text{(added and subtracted } f(\boldsymbol{x}_i)\text{)}$$

$$= E_{D,\epsilon}\left(\sum_{i=1}^{N}\left[\epsilon_i + f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)$$

$$= \sum_{i=1}^{N}\left[\overbrace{\underbrace{E_\epsilon(\epsilon_i^2)}_{=\sigma_{\epsilon_i}^2}}^{\text{by equation (2)}} + E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right) + 2\overbrace{\underbrace{E_\epsilon(\epsilon_i)}_{=0}}^{\text{by equation (2)}} E_D\left(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right] \text{(lin. prop. of } E$$

$$= \sum_{i=1}^{N}\left[\sigma_{\epsilon_i}^2 + E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)\right] = \text{Var}(\epsilon) + \sum_{i=1}^{N}\left[E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)\right] \qquad (5)$$

Calculations explicitly done by the author for educational purposes.

Now let me explain the meaning of each line in equation (5). In the first line, I calculate the expectation value of the cost function of the test dataset *D,* where in the first equality I wrote the cost function explicitly. In the second line, in equation (5), I added and subtracted the function ***f(x)*** at a given value of ***x*** and used equation (2) where I wrote $\epsilon = y\text{-}f$. In the fourth equality line, first I expanded the quadratic expression, and second I used the linear and product properties of the expectation value *E* for random variables. After that, I still used equation (2) to calculate the variance, and mean of the components of error ϵ, where its mean is zero for a random, normally distributed and uncorrelated error components. In the last line of equation (5), I used the fact that the sum of each error variance component gives the ***total error variance***, Var(ϵ), or just the ***noise***. However, it still remains to calculate the last term in the last line in equation (5).

If you managed to follow me so far in all steps of equation (5), then I must congratulate you again. Now, as I mentioned above, it remains to calculate the last term in the last line in equation (5) which is:

$$\sum_{i=1}^{N}\left[E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i)-f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)\right]$$

One important thing to note in the just above expression is that I am taking the expectation value on possible different datasets *D* and error instances ϵ. In fact, as I mentioned when derived equation (5), the calculation of the BV error presented here is the most general one because it takes into account different error instances and different datasets, therefore it is a *generalised BV error*. Now I can write the last term in the last line in equation (5) as follows:

$$E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i)-f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)=E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i)-\overbrace{E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)+E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)}^{=0}-f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)$$

$$=E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2\right)+E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2\right)+$$

$$2E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]\left[f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]\right)\qquad\text{(expanded the quadratic form}$$

$$=\underbrace{\left[f(\boldsymbol{x}_i)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2}_{\text{independent of }D}+E_D\left(\left[f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2\right)+$$

$$2\underbrace{\left[f(\boldsymbol{x}_i)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]}_{\text{independent of }D}\underbrace{E_D\left[f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]}_{=0}$$

$$=\left[f(\boldsymbol{x}_i)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2+E_D\left(\left[f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)-E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2\right)\qquad(6)$$

Calculations explicitly done by the author for educational purposes.

In the first line in equation (6), I added and subtracted a term that sum equals zero. In the second line of (6), I expanded the quadratic form and then used the linearity property of the expectation value *E* on each term. In the third line of (6), the first term is just a number and its expectation value is the number itself and is independent of *D,* the second term depends on *D,* while the third term is equal to zero because of the general property *E(X-E(X))=E(X)-E(X)=0* of a generic random variable *X*. In the fourth line of *(6)* is the remaining expression after all manipulations.

Now I use the last line in equation (6) into the last line in equation (5), and get the final result:

$$E_{D,\epsilon}(C_{\text{test}}) = \text{Var}(\epsilon) + \sum_{i=1}^{N}\left(E_{D,\epsilon}\left(\left[f(\boldsymbol{x}_i) - f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right]^2\right)\right)$$

$$= \text{Var}(\epsilon) + \underbrace{\sum_{i=1}^{N}\left[f(\boldsymbol{x}_i) - E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2}_{\text{Bias}^2[f(\boldsymbol{x};\bar{\boldsymbol{\theta}}_D)]} + \underbrace{\sum_{i=1}^{N} E_D\left(\left[f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D) - E_D\left(f(\boldsymbol{x}_i;\bar{\boldsymbol{\theta}}_D)\right)\right]^2\right)}_{\text{Var}(f(\boldsymbol{x};\bar{\boldsymbol{\theta}}_D))}$$

$$= \text{Var}(\epsilon) + \text{Bias}^2[(f(\boldsymbol{x};\bar{\boldsymbol{\theta}}_D)] + \text{Var}(f(\boldsymbol{x};\bar{\boldsymbol{\theta}}_D)) \quad (7)$$

Calculations explicitly done by the author for educational purposes.

---

5. Conclusions and remarks

Equation (7) is the final expression of our journey, where the expectation value of the test dataset cost function *C* is equal to the sum of ***total variance*** of the irreducible(or intrinsic) error ϵ, the ***total Bias*** squared and ***total variance*** of the learned approximation function.

Now that I derived equation (7), one might ask: what does it imply? what is the interpretation of equation (7)? ***The first point*** to note is that the test dataset output error is always bigger than the irreducible noise(Var(ϵ)), so the total noise after the learning process is always bigger than the initial noise.

The second point to note is that of the definition of Bias in (7), where the Bias takes into account the difference between the true function $f(\boldsymbol{x})$ at a given point to the learned function that depends on the learned parameter vector at the same point. If we are good at choosing the right form of the learned function, then its difference to the true model

function would be minimal and close to zero and the Bias error would be close to zero as well. So, this means that the Bias takes into account our accuracy in choosing the right function to model our data. If we pick a simpler linear function to model a given dataset where the true function is for example an exponential function, then our bias error would be large because our guess is very poor.

The third point to note is that of the definition of variance of the learned function in (7). One can observe that in the definition of variance, there is the difference between the learned function at a given point to the *expectation value* of the learned function over *D at the same point*. However, the expectation value depends on the dataset and on its *size*. So, the variance of the learned function gives the error that is generated due to the use of different datasets in our model and gives the difference between the learned function to its mean value calculated over different possible datasets.

The fourth and last point to note is that the BV error relation in equation (7) has been obtained by using the Euclidean metric for the test dataset cost function *C. If I had used a different metric, like for example the Manhattan distance metric, then the BV error relation would not necessarily be the same as that obtained in equation (7). Thus, the BV error relation depends on the metric used to calculate the distance.*

By Damian Ejlli on July 29, 2021.

Canonical link

Exported from Medium on October 8, 2021.