Foundation of Machine Learning (IT 582)

Autumn 2022

Pritam Anand

Bais-Variance Decomposition

Given the training set $\{(x_i, y_i) : x_i \in \mathbf{R}^n, y_i \in \mathbf{R} \text{ for } i = 1, 2, ..l, \}$, the regression problem attempts to estimate the relationship between the independent variable x and dependent variable y by choosing an appropriate function f from a given set of the function F. For unseen data point x_*, our desired estimate $f(x*)$ should approximate the y_* well.

Let us recall the basic assumptions about our regression models. The first assumption is that the data points (x_i, y_i) come from a fixed distribution D and are also independent. Also, we consider

$$y_i = f_0(x_i) + \epsilon_i, \tag{1}$$

where, $E(\epsilon_i) = 0$ and the target estimate of $f_0(x_i)$ is $E(y/x_i)$.

In our Least Squared methodology, we need to find a function $f(x)$ such that $E(y - f(x))^2$ is minimum , which requires the access of every (x_i, y_i) of D. In practice, the minimization of the $E(y - f(x))^2$ (Structural Risk) is difficult, as we usually have access to a sample T of population D. So, what maximum we can do is $\min_{f\in F} \sum_{i=1}^{l} (y_i - f(x_i))^2$ (Empirical Risk).

Let us suppose that using the information of sample T, we have estimated a function $f_T(x)$, then we hope that the $f_T(x)$ should generalize well on unseen data points. In our Least Square methodology, we hope that $E_T(y_i - f_T(x_i))^2$ should be the least as possible. Let us attempt to decompose the least square error obtained by the $f_T(x)$ on test data points.

$$E_T(y_i - f_T(x_i))^2 = E_T(y_i - f_0(x_i) + f_0(x_i) - f_T(x_i))^2$$
$$= E_T(y_i - f_0(x_i))^2 + E(f_0(x_i) - f_T(x_i))^2 + 2E_T\big((y_i - f_0(x_i))(f_0(x_i) - f_T(x_i))\big) \tag{2}$$

At first, we show that $E_T\big((y_i - f_0(x_i))(f_0(x_i) - f_T(x_i))\big) = 0$ as follows.

$E_T\big((y_i - f_0(x_i))(f_0(x_i) - f_T(x_i))\big) = E_T(y_i f_0(x_i)) - E_T(y_i f_T(x_i)) + E_T(f_0(x_i) f_0(x_i)) + E_T(f_0(x_i) f_T(x_i))$
$= E_T((f_0(x_i) + \epsilon_i) f_0(x_i)) - E_T((f_0(x_i) + \epsilon_i) f_T(x_i)) + E_T(f_0(x_i) f_0(x_i)) + E_T(f_0(x_i) f_T(x_i))$
, considering $\epsilon_i = y_i - f_0(x_i)$ from (1)
$= E_T(f_0(x_i) f_0(x_i)) + E_T(\epsilon_i) f_0(x_i)) - E_T(f_0(x_i) f_T(x_i)) - E_T((\epsilon_i) f_T(x_i)) + E_T(f_0(x_i) f_0(x_i)) + E_T(f_0(x_i) f_T(x_i))$
$= 0.$

It reduces the (2) as

$$
\begin{aligned}
E_T(y_i - f_T(x_i))^2 &= E_T(y_i - f_0(x_i))^2 + E(f_0(x_i) - f_T(x_i))^2 \\
&= E_T(\epsilon_i)^2 + E_T(f_0(x_i) - f_T(x_i))^2, \text{ considering } \ \epsilon_i = y_i - f_0(x_i) \ \text{ from (1)} \\
&= E_T(\epsilon_i)^2 + (E_T(f_0(x_i) - f_T(x_i)))^2 + \mathrm{Var}_T(f_0(x_i) - f_T(x_i)), \\
&\text{considering } \ E(Z^2) = (E(Z))^2 + \mathrm{Var}(Z) \\
&= E_T(\epsilon_i)^2 + (E_T(f_0(x_i) - f_T(x_i)))^2 + \mathrm{Var}_T(f_T(x_i)), \\
&\quad \text{considering } \ \mathrm{Var}(a - Z) = \mathrm{Var}(Z).
\end{aligned} \tag{3}
$$

The first term in (3), $E_T(\epsilon_i)^2$ is irreducible error. It depends upon the variance of noise in data. The term $(E_T(f_0(x_i) - f_T(x_i)))$ in (3) is *bias* that explains, how far is our estimated function from target function $f_0(x)$ on average. The third term in (3) is variance of estimates $f_T(x_i)$. Now, we can conclude that

$$E_T(y_i - f_T(x_i))^2 = \text{Irreducible Error} + \text{Bias}^2 + \text{Variance} \tag{4}$$

We can not reduce the error from $E_T(\epsilon_i)^2$. In our best case, we can obtain the estimate $f_T(x) = f_0(x) = E(y/x)$ but, still our estimate will obtain the least square error $E_T(\epsilon_i)^2$ on test data points. We can work on the variance and bias of our estimate for reducing its generalization error in the least squared sense.