# ML Notes

Somnath Sikdar

December 20, 2020

# Contents

# Chapter 1

# Bias-Variance Decomposition

The bias-variance decomposition is the decomposition of the generalization error of a learning algorithm into the sum of the bias, the variance and an irreducible error term.

Consider a setting where a response variable $Y$ is related to a set of predictor variables $X \in \mathbf{R}^p$ as follows: $Y = f(X) + \varepsilon$, where $f\colon \mathbf{R}^p \to \mathbf{R}$ is some deterministic function and $\varepsilon$ is white noise. That is, $\mathbf{E}\left[\varepsilon | X = x\right] = 0$ and $\mathrm{Var}(\varepsilon | X = x) = \sigma^2$, for some variance $\sigma^2$. Let us assume that there is an underlying data distribution $\mathscr{D}(X, Y)$ which is unknown to us. We have an algorithm that, given an $n$-sized training set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ sampled iid from this distribution $\mathscr{D}$, yields $\hat{f}_S$, which is an approximation to the true function $f$. We assume that the algorithm itself is deterministic: that is, given the same $n$-sized sample $S$, it produces the same output $\hat{f}_S$.

What is the expected generalization error of this algorithm? This is defined as:

$$\mathbf{E}_{S \sim \mathscr{D}^n, (X,Y) \sim \mathscr{D}} \left[ (Y - f_S\hat{(X)})^2 \right]. \tag{1.1}$$

This is the expected error when we input the random sample $S$ to our algorithm to obtain $\hat{f}_S$ and then evaluate it on the random point $(X, Y)$. We may write this expected error as:

$$\mathbf{E}_{S \sim \mathscr{D}^n, (X,Y) \sim \mathscr{D}} \left[ (Y - \hat{f}_S(X))^2 \right] = \int \mathbf{E}_{(X,Y) \sim \mathscr{D}} \left[ (Y - \hat{f}_S(X))^2 \mid S \right] \cdot \mathbf{Pr}_{\mathscr{D}^n} \left\{ S \right\} \mathrm{d}S. \tag{1.2}$$

We will first work with the expectation term on the right-hand side of the above equation. That is, we will assume that the sample $S$ is fixed and then calculate the expected error wrt the approximation $\hat{f}_S$. In order to implify our notation a bit, we will index the expectation operator with the random variable to indicate which distribution is being referred to.

Let us write $(Y - \hat{f}_S(X))^2$ as $(Y - f(X) + f(X) - \hat{f}_S(X))^2$. Expanding, we get:

$$\begin{aligned} \mathbf{E}_{(X,Y)} \left[ (Y - \hat{f}_S(X))^2 | S \right] = \mathbf{E}_{(X,Y)} &\left[ (f(X) - \hat{f}_S(X))^2 | S \right] + \\ \mathbf{E}_{(X,Y)} &\left[ (Y - f(X))^2 | S \right] + \\ \mathbf{E}_{(X,Y)} &\left[ 2(Y - f(X))(f(X) - \hat{f}_S(X)) | S \right] \end{aligned} \tag{1.3}$$

Consider the last term on the right-hand side. We claim that this is 0. Indeed, we may write $\mathbf{E}_{(X,Y)}\left[2(Y - f(X))(f(X) - \hat{f}_S(X))|S\right]$ as

$$
\begin{aligned}
&= \iint 2 \cdot (y - f(x)) \cdot (f(x) - \hat{f}_S(x)) \cdot p_{X,Y}(x,y) \mathrm{d}y \, \mathrm{d}x \\
&= \iint 2 \cdot (y - f(x)) \cdot (f(x) - \hat{f}_S(x)) \cdot p_{Y|X}(y|x) \cdot p_X(x) \mathrm{d}y \, \mathrm{d}x \\
&= \int 2 \cdot (f(x) - \hat{f}_S(x)) \cdot \mathbf{E}_{Y|X}\left[Y - f(x)|X = x\right] \cdot p_X(x) \mathrm{d}x.
\end{aligned}
$$

Now we know that $\mathbf{E}_{Y|X}\left[Y - f(x)|X = x\right] = \mathbf{E}_{Y|X}\left[\varepsilon|X = x\right] = 0$. Hence this whole expression evaluates to 0 as claimed.

Next consider the second term $\mathbf{E}_{(X,Y)}\left[(Y - f(X))^2|S\right]$ of Equation (1.3). Since the term $(Y - f(X))^2$ does not depend on the sample $S$ that is chosen, this further simplifies to $\mathbf{E}_{(X,Y)}\left[(Y - f(X))^2\right]$. This is simply the variance of the error term and is equal to $\sigma^2$. Finally, consider the first term of Equation (1.3) which is $\mathbf{E}_{(X,Y)}\left[(f(X) - \hat{f}_S(X))^2|S\right]$. We use the trick of adding and subtracting as before. This time around, we add and subtract the term $g(X) \coloneqq \mathbf{E}_S\left[\hat{f}_S(X)\right]$ to obtain:

$$
\begin{aligned}
(f(X) - g(X) + g(X) - \hat{f}_S(X))^2 = (f(X) - g(X))^2 + (g(X) - \hat{f}_S(X))^2 + \\
2 \cdot (f(X) - g(X)) \cdot (g(X) - \hat{f}_S(X)).
\end{aligned} \tag{1.4}
$$

We will evaluate each of these terms by directly plugging them in the expression in Equation (1.2). Let's examine the first term $(f(X) - g(X))^2$. Plugging this in the said expression, we see that we have to evaluate:

$$
\begin{aligned}
\int \mathbf{E}_{(X,Y)}\left[(f(X) - g(X))^2|S\right] \mathbf{Pr}_{S\sim\mathscr{D}^n}\left\{S\right\} \mathrm{d}S &= \mathbf{E}_{(X,Y)}\left[(f(X) - g(X))^2\right] \int \mathbf{Pr}_{S\sim\mathscr{D}^n}\left\{S\right\} \mathrm{d}S \\
&= \mathbf{E}_{(X,Y)}\left[(f(X) - g(X))^2\right].
\end{aligned} \tag{1.5}
$$

The first equality holds because neither $f(X)$ nor $g(X)$ depends on the sample $S$ chosen. The resulting expression is the expected squared bias of the estimator $\hat{f}_S$.

Forging ahead, we evaluate the next term which is $(g(X) - \hat{f}_S(X))^2$.

$$
\int \mathbf{E}_{(X,Y)}\left[(g(X) - \hat{f}_S(X))^2|S\right] \mathbf{Pr}_{S\sim\mathscr{D}^n}\left\{S\right\} \mathrm{d}S = \mathbf{E}_{S,(X,Y)}\left[(g(X) - \hat{f}_S(X))^2\right]. \tag{1.6}
$$

This term is the variance of the estimator $\hat{f}_S$ obtained by our algorithm.

Finally, we evaluate the term $2 \cdot (f(X) - g(X)) \cdot (g(X) - \hat{f}_S(X))$. We now have to evaluate this integral:

$$
\int \mathbf{E}_{(X,Y)}\left[2 \cdot (f(X) - g(X)) \cdot (g(X) - \hat{f}_S(X))|S\right] \mathbf{Pr}_{S\sim\mathscr{D}^n}\left\{S\right\} \mathrm{d}S \tag{1.7}
$$

Recall that this is just the expectation wrt the distribution of the random sample $S$ *and* the random data point $(X, Y)$ at which we want to evaluate the predictor $\hat{f}_S$ output by our algorithm. Instead of first conditioning on $S$, we can condition on the random data point $(X, Y)$ and write:

$$\iint \mathbf{E}_S \left[2 \cdot (f(x) - g(x)) \cdot (g(x) - \hat{f}_S(x)) | X = x, Y = y\right] p_{X,Y}(x, y) \mathrm{d}y \, \mathrm{d}x. \tag{1.8}$$

Now the term $f(x) - g(x)$ does not depend on $S$ and $\mathbf{E}_S \left[g(x) - \hat{f}_S(x)\right] = 0$ and hence the above integral evaluates to 0.

To summarize, we may write the expected generalization error of our algorithm as:

$$\begin{aligned} \mathbf{E}_{S \sim \mathscr{D}^n, (X,Y) \sim \mathscr{D}} \left[(Y - \hat{f}_S(X))^2\right] = & \underbrace{\mathbf{E}_{(X,Y)} \left[(\mathbf{E}_S \left[\hat{f}_S(X)\right] - f(X))^2\right]}_{\text{expected squared bias}} \\ & + \underbrace{\mathbf{E}_{S,(X,Y)} \left[(\hat{f}_S(X) - \mathbf{E}_S \left[\hat{f}_S(X)\right])^2\right]}_{\text{variance}} \\ & + \underbrace{\sigma^2}_{\text{irreducible error}} . \end{aligned}$$