ML Fundamentals: Learning Theory

February 2, 2021

1 Learning Theory

In the below figure 1, we can see how bias is the cause of under-fitting and variance is the cause of over-fitting the data.

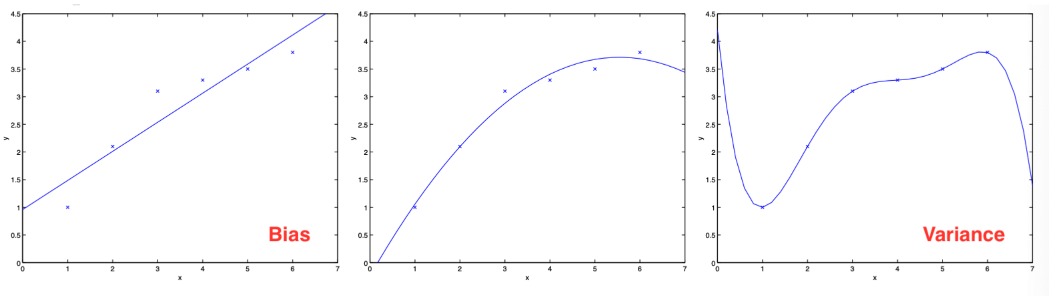

Figure 1: Illustration Bias and Variance on Fitting the Data

Parameters view for Bias Variance tradeoff can be seen in the figure 2, where center of the space is true parameter.

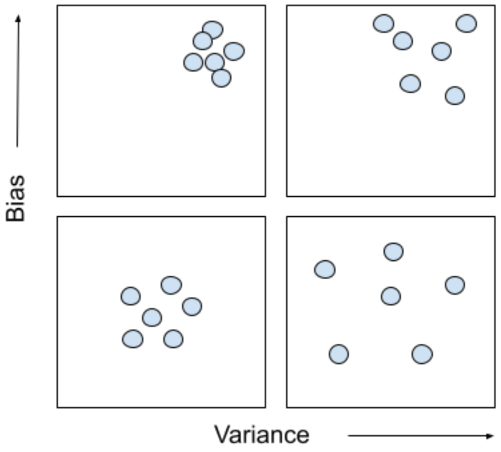

Figure 2: Parameters view of Bias and Variance

How bias and variance impact the generalisation error, adding more samples helps or making hypothesis class bigger by increasing model complexity helps? What are the bounds of number of parameters and number of possible hypothesis on the generalisation errors? Lets try to answer these questions.
We would use two lemmas to prove important results on learning theory.
Lemma 1: Union Bound. Let $A_1, A_2, .., A_k$ be k different events. Then,

$$P\left(A_1 \cup \cdots \cup A_k\right) \leq P\left(A_1\right) + \ldots + P\left(A_k\right)$$

Lemma 2: Hoeffding inequality, Let $Z_1, Z_2, .., Z_m$ be m independent and identically distributed (iid) random variables drawn from Bernoulli(ϕ) distribution, i.e. $P(Z_i = 1) = \phi$, and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} =$

$(1/m) \sum_{i=1}^{m} Z_i$ be the estimated Bernoulli parameter. Then,

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp\left(-2\gamma^2 m\right)$$

Above lemma 2 is also called **Chernoff bound** in learning theory.
Now lets say, training error of a hypothesis h on training data of m samples is,

$$\hat{\varepsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\left\{h\left(x^{(i)}\right) \neq y^{(i)}\right\}$$

And, generalisation error would be,

$$\varepsilon(h) = P_{(x,y) \sim \mathcal{D}}(h(x) \neq y)$$

Two assumptions, training and testing data is drawn from the same distribution, and each sample is independently drawn, are called **PAC (Probably Approximately Correct) assumptions.**
Empirical Risk Minimization (ERM) is a process to choose the parameters (or hypothesis h) from a set of possible parameters (or hypothesis class $\mathcal{H}$), which as minimum error on training data. It can be written as,

$$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}\left(h_\theta\right)$$

Or, in a broader way,

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

1.1 The case of finite H

Lets consider a Bernoulli random variable, $Z = 1\left\{h_i(x) \neq y\right\}$, where $(x, y) \sim \mathcal{D}$. The estimated error for the hypothesis h would be,

$$\hat{\varepsilon}\left(h_i\right) = \frac{1}{m} \sum_{j=1}^{m} Z_j$$

By applying **Chernoff bound**, we can establish relation between estimated error or generalised error as follow,

$$P\left(\left|\varepsilon\left(h_i\right) - \hat{\varepsilon}\left(h_i\right)\right| > \gamma\right) \leq 2 \exp\left(-2\gamma^2 m\right)$$

This says, that, for a particular hypothesis h_i, training error would be close to generalisation error with high probability when m is large. But how about establishing the relation from a true minimum generalisation error?

Let say we have k hypothesis in class $\mathcal{H}$. We can apply **union bound** as follow:

$$\begin{aligned}
P\left(\exists h \in \mathcal{H}. \left|\varepsilon\left(h_i\right) - \hat{\varepsilon}\left(h_i\right)\right| > \gamma\right) &= P\left(A_1 \cup \cdots \cup A_k\right) \\
&\leq \sum_{i=1}^{k} P\left(A_i\right) \\
&\leq \sum_{i=1}^{k} 2 \exp\left(-2\gamma^2 m\right) \\
&= 2k \exp\left(-2\gamma^2 m\right)
\end{aligned}$$

If we subtract both sides from 1, we find that

$$\begin{aligned}
P\left(\neg \exists h \in \mathcal{H}. \left|\varepsilon\left(h_i\right) - \hat{\varepsilon}\left(h_i\right)\right| > \gamma\right) &= P\left(\forall h \in \mathcal{H}. \left|\varepsilon\left(h_i\right) - \hat{\varepsilon}\left(h_i\right)\right| \leq \gamma\right) \\
&\geq 1 - 2k \exp\left(-2\gamma^2 m\right)
\end{aligned}$$

Above derivation is called **uniform convergence**, which states that "there exist at-least one hypothesis, for which training error is with in γ margin of the generalisation error", with the probability of $1 - \delta$. Where, $\delta = 2k \exp\left(-2\gamma^2 m\right)$. So, we can ask questions like, how many samples would be require to satisfy this probability bound $1 - \delta$?

$$m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

Which states (also known as **sample complexity**), to assure higher probability (lower δ) of such hypothesis where training error is in bound of generalisation error, we would require higher m.

Bigger hypothesis class $\mathcal{H}$, would have higher k, and which would be the case for more complex models, and for which, higher m would be required.
Stricter margins γ would require higher m.

How about a bound on training error with respect to the best possible hypothesis in $\mathcal{H}$, .e. h^*. ERM would select $\hat{h}$ as the training error of $\hat{h}$ would be lesser than training error of h^*. We can state,

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}\left(h^*\right) + \gamma \\ &\leq \varepsilon\left(h^*\right) + 2\gamma \end{aligned}$$

Let's put all this together into a theorem.
Theorem. Let $|\mathcal{H}| = k$, and let any m, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$