3 Underlying Machine Learning Theory

3.1 Statistical Learning theory

Ideally, for all machine learning methods we minimize the empirical risk - hoping that it is a good surrogate for the real risk. Ultimately, one would like to get a strategy that generalizes best from training to testing data (By testing I mean, unseen independent data). That is, we do not want the model to overfit i.e. fit too well that it fails on new unseen data.

Statistical learning theory provides insights to these "over fitting"/"Generalization" phenomena in Machine learning by theorizing these concepts and by proposing solutions. In fact, there are three essential statistical learning metrics that determine Overfitting: On the first hand the "Generalization error" denoted $||\hat{r}(\hat{s}) - r(\hat{s})||_p$ for some strategy $\hat{s}$ and for some norm p, measures the difference between the in-sample and the out of sample loss for some strategy evaluated in-sample. On the other hand, statistical learning theory, defines two other metrics : The estimation and the optimization errors through risk decomposition. In fact, any risk associated to some strategy can be decomposed by simple algebraic manipulation. We can thus write the the "Estimation / Approximation" decomposition :

$$\underbrace{r\left(\hat{s}_{\text{erm}}\right) - r_0}_{\text{Excess risk}} = \underbrace{r\left(\hat{s}_{\text{erm}}\right) - r\left(s^*\right)}_{\substack{\text{Estimation} \\ \text{error}}} + \underbrace{r\left(s^*\right) - r_0}_{\substack{\text{Approximation} \\ \text{error}}} \tag{2}$$

With $\hat{s}_{\text{erm}}$ [17]the estimated strategy that empirically minimizes the loss on training data such that $\hat{s}_{erm} = \text{Arginf}_{s \in S}\, \hat{r}(s)$; s^* the strategy that minimizes the loss on unseen data for some hypothesis space $\mathbb{S}$ such that $s^* = \text{Arginf}_{s \in S_{all}}\, \hat{r}(s)$ and s_0 the strategy that minimizes the loss on unseen data on the whole hypothesis space such that:$s_0 = \text{Arginf}_{s \in S_{all}}\, r(s)$.

Note that the risk function $r(.)$ we have used in 2 is the true risk [18] - it is the loss evaluated on unseen data- and the decomposition formalizes the idea that learning is essentially about picking some strategy s belonging to some set of strategies $\mathbb{S}$ which is , in turn , a subset of the whole space of strategies: Explicitly, one can write : $s \in \mathbb{S} \subset \mathbb{S}_{all}$.[19]

Having presented these metrics; the generalization error [20] , the approximation error[21] and the estimation error [22] our goal is to ultimately minimize all three of them for the empirically minimized

[17] "Erm" is an abbreviation for Empirical Risk Minimization

[18] Not $\hat{r}(.)$ - The empirical one

[19] For example, if one chooses to modelize the data with a second order degree polynomial regression function and hence end up with some ERM fit: s is represents the ERM fit, which is contained in $\mathbb{S}$, the $P_2(x)$ polynomial, which in turn is contained in $\mathbb{S}_{all}$, the large hypothesis space of all $P_N(x) : n > 2$.

[20] The distance between model's risk evaluated on some sample and the expected value of the risk

[21] i.e. How far is the best model in a chosen hypothesis class from the true best hypothesis

[22] i.e. How far is the best model to the estimated model in some chosen hypothesis space

strategy to serve as an effective proxy to the true strategy. We thus want to study the dynamics of all three of them. From the one hand, it is easy to deduce that, the approximation error, $(r(s^*) - r_0)$ is inversely related to the complexity: That is because for $s_1^* \in S_1 \& s_2^* \in S_2$ with $S_1 \subset S_2$ (i.e. S_1 belongs to a more complex hypothesis class), $r\left(s_1^*\right) \geq r\left(s_2^*\right)$, i.e. the optimal risk can only reduce or stay the same when exposed to new strategies.Hence, we minimize the approximation error by increasing model's complexity.

However, understanding the dynamics of both the generalization error and the estimation is not trivial. Unlike the approximation error, it is not clear how a broader class of hypothesis would affect both errors (We can individually study the dynamics of some risk associated to some strategy - this was done for the approximation error whereby only one risk matters, the other , r_o being fixed. However, studying the dynamics of the difference between the 2 varying risks is not trivial).

Classical statistical learning theory (Vapnik [29]) proposes to bound both errors measures in order to study their dynamics. The rationale behind the derivation of these bounds is the following: Probability theory offers many methods to derive bounds on deviations from expectations [23], which we express as $P(|x - \mathbb{E}(x)| \geq \varepsilon) \leq \alpha_x(n, \varepsilon)$ with α depending on the probabilistic distribution of X. We can generalize this to risk functions: $P(|\hat{r}(s) - r(s)| \geq \varepsilon) \leq \alpha_{\hat{r}(s)}(n, \varepsilon)$. The issue here is that the bound depends on a specific value of s. Accordingly, statistical learning theory derives bounds on $Max_{s \in S}|\hat{r}(s) - r(s)|$ (called "uniform bounds") thus getting $P\left(\max_{s \in S}|\hat{r}(s) - r(s)| \geq \varepsilon\right) \leq \alpha(n, \varepsilon)$; i.e. a bound that is not specific to one strategy. This deviation bounds both the generalization and the estimation error. Knowing α (i.e. studying the bound of this derivation) gives us an indication about the dynamics of both the estimation and the generalization error.

$$\text{Estimation error} \leq Max|\hat{r}(s) - r(s)| \leq \text{Some bound}$$

$$\text{Generalization error} \leq Max|\hat{r}(s) - r(s)| \leqslant \text{Some Bound}$$

Statistical learning theory derives the bounds with respect to different complexity measures. Notable examples are bounds with respect to $|H|$, the size of the hypothesis space (But this is not convenient as $|S| \longrightarrow \infty$ usually), the Vapnik–Chervonenkis dimension or even the Rademacher complexity (Mohri et al. 2012 [18]). Each bound has its own specificity (There are data dependent bounds, others are distribution dependent etc ...) but the general idea is that both the generalized and the estimation error are upper bounded by some measure of model's complexity.

Thus, from one hand the approximate error is minimized by reducing model's complexity, and from the other hand, the complexity limits how bad the estimation and the generalization errors can go.

[23] Starting from Markov's inequality, one can derive, under particular conditions, many different bounds such as Chebychev , exponential Markov (Chernoff) bounds etc...

Here, complexity is positively proportional to the bounds of the error. This result is very important as it establishes a theoretical "interpretation" of the estimation/approximation trade off [24],as well as the variance bias trade off .

These derived bounds constitute the theory that supports the dynamics of the approximation / estimation / bias / variance errors with respect to complexity: The models must be complex enough to get a low approximation error (or "bias" in regression terms) but not too complex that the complexity measure would increase the limiting upper bounds of the generalization and estimation error (or Variance). Notice here that the bounds constitute theoretical support for the Occam razor principal; i.e. the principal of parsimony that posits that the model should be as simple as possible.

The Variance-Bias trade off was not discussed in this paper, but its rationale and derivation method is similar to that of the estimation approximation error, yet , those are two different concepts. They are both derived by decomposition. The Variance bias decomposition is derived by breaking down the sample risk such that $\mathbb{E}[R(\hat{f}))] = \underbrace{\mathbb{E}_{xy}\left((y - \mathbb{E}_{y,x}(y))^2\right)}_{\text{Intrinsic Noise}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_D\left(\hat{f}_{(x)}\right) - \mathbb{E}_{y|x}(y)\right]^2}_{\text{Variance}} + \underbrace{\mathbb{E}_x\left[\mathbb{E}_D\left(\hat{f}_{(x)} - \mathbb{E}_D\left(\hat{f}_{(x)}\right)\right)^2\right]}_{\text{Bias}}$ (German et al. 1992). However, the variance bias decomposition is not equivalent to the Estimation/Approximation decomposition, further algebraic manipulations actually proves it (Refer to the variance bias decomposition section in appendix for a clear illustration of the difference between both decompositions).

This presented rationale derived from statistical learning theory provides the underlying theoretical background of the machine learning methods that I present.

[24] I purposely did not say "proof", but "interpretation", because upper bounds alone do not guarantee the trade off