Chapter 5

Motivation

Previously, we have seen that overfitting can occur. To overcome this problem, we restrict the search space to some hypothesis class H. This can be seen as incorporating some *prior* knowledge that the learner has about the task. There is the question of whether such prior knowledge is needed to be successful at learning. Could there be a universal learner who has no prior knowledge about a learning task and can succeed at any task?

Specifically, suppose we have a learning task defined by some unknown distribution D over $X \times Y$, where the goal of the learner is to find a predictor h, whose risk is small enough. The question is therefore whether there exists a learning algorithm A and a training set of size m, such that for every distribution D, if A receives m i.i.d. examples from D, there is a high chance it outputs a predictor h that has a low risk.

The **No-Free-Lunch** theorem states that no such universal learner exists. To be more precise, the theorem states that for binary classification prediction tasks, for every learner there exists a distribution on which it fails. We say that the learner fails if, upon receiving i.i.d. examples form that distribution, its output hypothesis is likely to have a large risk, whereas for the same distribution, there exists another learner that will output a hypothesis with a small risk. In other words, the theorem states that no learner can succeed on all learnable tasks - every learner has tasks on which it fails while other learners succeed.

The No-Free-Lunch Theorem

Theorem 5.1 (No-Free-Lunch)

> Let A be any learning algorithm for the task of binary classification with respect to the 0-1 loss over a domain X. Let m be any number smaller than $|X|/2$, representing a training set size. Then, there exists a distribution D over $X \times \{0, 1\}$ such that:
>
> 1. There exists a function $f : X \rightarrow \{0, 1\}$ with $L_D(f) = 0$.
> 2. With probability of at least 1/7 over the choice of $S \sim D^m$ we have that $L_D(A(S)) \geq 1/8$.

This theorem states that for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner.

> Let C be a subset of X, $|C| = 2m$.
>
> Note that there are $T = 2^{|C|} = 2^{2m}$ possible functions from C to $\{0, 1\}$. Denote these functions by $f_1, \dots, f_T$. For each such function, let D_i be a distribution over $C \times \{0, 1\}$ defined by

> $$D_i(\{(x,y)\}) = \begin{cases} 1/|C| \text{ if } y = f_i(x) \\ 0 \text{ otherwise} \end{cases}$$
>
> So for some pair (x, y), the probability it is chosen is $1/|C|$ if the label y is really the true label according to f_i.
>
> We will show that for every algorithm, A, that receives a training set of m examples from $C \times \{0, 1\}$ and returns a function $A(S) : C \to \{0, 1\}$, it holds that the worst case expected loss is greater than or equal to 1/4:
>
> $$\max_{i \in [T]} \mathbb{E}_{S \sim D_i}[L_{D_i}(A(S))] \geq 1/4$$
>
> Then $\Pr[L_D(A'(S)) \geq 1/8] \geq 1/7$, which is what we need to prove.

Intuition: Any learner which observes only half of the instances in C has no information on what should be the labels on the rest of instances in C, if all functions are possible. Therefore, there exists some target function f, that would contradict the labels that $A(S)$ predicts on the unobserved instances in C.

Corollary 5.2

> Let X be an infinite domain set and let H be the set of all function from X to $\{0, 1\}$. Then, H is not PAC learnable.

How can we prevent such failures? We should use our prior knowledge about a specific learning task to avoid the distributions that will cause us to fail when learning the task.

One way to do this is by restricting our hypothesis class.

For example, we can choose a class which includes the hypothesis with 0 error. But, as we just saw, we cannot just choose the most expressive class - the class of all functions over the given domain. This tradeoff is the bias-complexity tradeoff.

Error Decomposition

Deompose the error of an ERM_H predictor into 2 components as follows. Let h_S be an ERM_H hypothesis. Then:

$$L_D(h_S) = \epsilon_{app} + \epsilon_{est} = \min_{h \in H} L_D(h) + \epsilon_{est}$$

The Approximation Error

- This is the minimum risk achievable by a predictor in the hypothesis class. It measures how much risk we have because we restrict ourselves to a specific class, which is how much inductive bias we have.
- Does not depend on the sample size

- Determined by the hypothesis class chosen
- Intuitively, how close we can get to the “true” labelling function, which is of course affected by how well our class can approximate the true function

The Estimation Error

- The difference between the approximation error and the error achieved by the ERM predictor.
- This error results be cause the empirical risk is only an estimate of the true risk (finite sample size)
- The quality of this estimation depends on the training set size and on the size, or complexity, of the hypothesis class.
- $\uparrow \propto \log(|H|)$
- $\downarrow \propto m$

Bias-Complexity Tradeoff

Since our aim is to minimise the total risk, we face a tradeoff. If we choose a very rich class, the approximation error decreases, but this might inrease the estimation error, as a rich class might lead to **overfitting**.

On the other hand, choosing a limited class reduces the estimation error but might increase the approximation error, cuasing **underfitting**.