Chapter 6

Motivation

In the previous chapter, we decomposed the error of the ERM_H rule into approximation error and estimation error. The approximation error depends on the fit of our prior knowledge (as reflected by the choice of the hypothesis class $\mathcal{H}$) to the underlying unknown distribution. In contrast, the definition of PAC learnability requires that the estimation error would be bounded uniformly over all distributions.

Our current goal is to figure out which classes H are PAC learnable, and to characterise exactly the sample complexity of a given hypothesis class.

Indeed, infinite classes can be learnable, and thus, finiteness of the hypothesis class is not a necessary condition for learnability.

The VC-Dimension

While finiteness of H is a sufficient condition for learnability, it is not a necessary condition.

Definition 6.2 (Restriction of H to C)

> Let H be a class of functions from $\mathcal{X}$ to {0,1} and let $C = \{c_1, \dots, c_m\} \subset \mathcal{X}$. The restriction of H to C is the set of functions from C to {0,1} that can be derived from H. That is,
>
> $$H_C = \{(h(c_1), \dots, h(c_m)) : h \in H\}$$
>
> where we represent each function from C to {0,1} as a vector in $\{0, 1\}^{|C|}$

```
Example
|C| = 3
(0, 0, 0)
(0, 0, 1)
...
(1, 1, 0)
(1, 1, 1)

Then, H shatters C if we can find functions in H which exactly classify the poi
```

If the restriction of H to C is the set of all functions from C to $\{0, 1\}$, then we say that H shatters the set C.

Definition 6.3 (Shattering)

> A hypothesis class H shatters a finite set $C \subset \mathcal{X}$ if the restriction of H to C is the set of all functions from C to {0,1}. That is, $|H_C| = 2^{|}C|$.

By the No-Free-Lunch theorem, if H shatters some set C of size $2m$ then we cannot learn H using m examples. Intuitively, if a set C is shattered by H, and we receive a sample containing half the instances of C, the labels of these instances give us no information about the labels of the rest of the instances in C - every possible labeling of the rest of the instances can be explained by some hypothesis in H.

> If someone can explain every phenomenon, his explanations are worthless.

Definition 6.5 (VC-dimension)

> The VC-dimension of a hypothesis class H, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by H. If H can shatter sets of arbitrarily large size we say that H has inifinite VC-dimension.

Theorem 6.6

> Let H be a class of infinite VC-dimension. Then, H is not PAC learnable.

Note that, to show that VCdim(H) = d, we need to show that 1. There exists a set C of size d that is shattered by H. 2. Every set C o size d+1 is not shattered by H.

Theorem 6.7 (The Fundamental Theorem of Statistical Learning)

> Let H be a hypothesis class of functions from a domain $\mathcal{X}$ to {0,1} and let the loss function be the 0-1 loss. Then, the following are equivalent: 1. H has the uniform convergence property 2. Any ERM rule is a successful agnostic PAC learner for H 3. H is agnostic PAC learnable 4. H is PAC learnable 5. Any ERM rule is a successful PAC learner for H 6. H has a finite VC-dimension