Chapter 4

4.1 Uniform Convergence

Motivation: So far, the learner can output a hypothesis which minimises the empirical risk with respect to S. We hope that this hypothesis also minimises the error with respect to the true probability distribution. In order to do so, we would like to ensure that the empirical risks of all members of H are good approximations of their true risk.

In other words, we want that uniformly over all hypotheses in the hypothesis class, the empirical risk will be close to the true risk.

Definition 4.1 (ϵ-representative sample)

> A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class H, loss function ℓ, and distribution D) if
>
> $$\forall h \in H : |L_S(h) - L_D(h)| \leq \epsilon$$

In plain English: a sample is ϵ-representative if the absolute difference between the empirical and true risk is bounded by ϵ.

When we have a sample which is $\epsilon/2$-representive, the ERM learning rule is guaranteed to return a good hypothesis:

Lemma 4.2

> Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class H, loss function ℓ, and distribution D). Then, any output of $ERM_H(S)$ satisfies
>
> $$L_D(h_S) \leq \min_{h \in H} L_D(h) + \epsilon$$

Proof:

> 1. The empirical risk for the ERM hypothesis is bounded, by definition 4.1
>
> $$L_d(h_S) \leq L_S(h_S) + \epsilon/2$$
>
> 2. The empirical risk for any other hypothesis is greater than or equal to the ERM hypothesis, h_S .
>
> $$L_S(h_S) + \epsilon/2 \leq L_S(h) + \epsilon/2$$

3. The empirical risk for any other hypothesis is bounded, by definition 4.1

$$L_S(h) + \epsilon/2 \leq (L_d(h_S) + \epsilon/2) + \epsilon/2$$

Lemma 4.2 implies that to ensure that the ERM rule is an agnostic PAC learner, we just need to show that with probability of at least $1 - \delta$ over the random choice of a training set, it will be an ϵ-representative training set.

Definition 4.3 (Uniform Convergence)

> A hypothesis class H has the **uniform convergence** property (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_H^{UC} : (0,1)^2 \to \mathbb{N}$ such that $\forall \epsilon, \delta \in (0,1)$ and for every probability distribution D over Z, if S is a sample of $m \geq m_H^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to D, then, with probability of at least $1 - \delta$, S is ϵ-representative.

Intuition:

A class with the uniform convergence property implies the existence of a function, m_H^{UC} such that for every ϵ, δ, with a large enough sample size (as defined by the function), we can do PAC learning.

So the function m_H^{UC} measures the minimal sample complexity of obtaining the uniform convergence property, which is the *number of examples needed* to ensure that with probability of at least $1 - \delta$ the sample would be ϵ-representative.

The term *uniform* refers to having a fixed sample size that works for all members of H and over all possible probability distributions over the domain.

Corollary 4.4

> If a class H has the uniform convergence property with a function m_H^{UC} then the class is agnostically PAC learnable with the sample complexity $m_H(\epsilon, \delta) \leq m_H^{UC}(\epsilon/2, \delta)$

With Corollary 4.4, we can conclude that finite classes are agnostic PAC learnable as long as uniform convergence holds for a finite hypothesis class.