

Chapter 2

Introduction to Learning Theory

In logic, the process of finding a general rule from a finite set of observations is called induction (Genesereth and Nilsson 1987, Chap. 7, pp. 161–176)¹. In machine learning, this inductive framework has been implemented according to the empirical risk minimization (ERM) principle, and its statistical properties have been studied in the theory developed by Vapnik (1999). The striking result of this theory is an upper bound of the generalization error of the learned function, expressed in terms of its empirical error on a giving training set and the complexity of the class of functions in use. This complexity reflects the ability or capacity of the function class to solve the prediction problem and is as large as it contains functions able to predict all possible outputs of observations in the training set. In other terms, the higher the capacity, the lower the empirical risk and the less we are guaranteed to achieve the main goal of the learning that is to have a low generalization error. Hence, we see that there is a compromise between the empirical error and the capacity of the function class, and that the aim of learning is to minimize the empirical error while controlling the capacity of the function class. This principle is called structural risk minimization and along with the ERM principle they are at the origins of many learning algorithms. Further, they also explain the behaviours of algorithms designed before the establishment of the learning theory by Vapnik (1999). The remainder of this chapter is devoted to a more formal presentation of these concepts under the binary classification setting which constitutes the initial part of the development of this theory.

2.1 Empirical Risk Minimization

In this section, we present the ERM principle by first presenting the notations that will be further used.

¹ The opposite reasoning, called deduction, is based on axioms and consists in producing specific rules (which are always true) as consequences of the general law.

2.1.1 Assumption and Definitions

We assume that the observations are represented in a vector space of fixed dimension d , $\mathcal{X} \subseteq \mathbb{R}^d$. The desired output of observations are assumed to belong to the set $\mathcal{Y} \subset \mathbb{R}$. Until the early 2000s, classification and regression were the two main frameworks developed in supervised learning. In classification, the output space $\mathcal{Y}$ is a discrete set and the prediction function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is called a classifier. When $\mathcal{Y}$ is a continuous interval of the real set, f is said to be a regression function. In Chap. 4, we present a new framework called, learning to rank that has recently been developed in both machine learning and information retrieval communities. A pair $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ thus denotes a labeled example and $S = (\mathbf{x}_i, y_i)_{i=1}^m \in (\mathcal{X} \times \mathcal{Y})^m$ indicates a training set. In the particular case of binary classification that we consider in this chapter, we note the output space by $\mathcal{Y} = \{-1, +1\}$ and a pair $(\mathbf{x}, +1)$ (respectively $(\mathbf{x}, -1)$) is called a positive (respectively negative) example.

The fundamental assumption of the learning theory is that examples are independently and identically distributed (i. i. d.) according to a fixed, but unknown, probability distribution $\mathcal{D}$. The identically distributed hypothesis ensures that observations are stationary while the independently distributed hypothesis states that each individual example provides a maximum information to solve the prediction problem. According to this hypothesis for a given prediction problem, examples of different training and test sets are all supposed to have the same behaviour.

Hence, this assumption characterizes the notion of representativeness of training and test sets with respect to the prediction problem, i.e. the training and unseen examples are assumed to be generated from a fixed source of information.

Another basic concept in machine learning is the notion of loss, also referred by risk or cost. For a given prediction function f , the disagreement between the desired output y of an example $\mathbf{x}$ and the prediction $f(\mathbf{x})$ is measured by an instantaneous loss function defined as:

$$e : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$$

Generally this function is a distance in the output space $\mathcal{Y}$ measuring the difference between the desired and the predicted outputs for a given observation. In the regression framework, the usual instantaneous loss functions are the ℓ_1 and ℓ_2 norms of the difference between both the desired and the predicted outputs of an observation. In binary classification the most common instantaneous loss considered is the 0/1 loss, which for an exemple $(\mathbf{x}, y)$ and a prediction function f is defined as

$$e(f(\mathbf{x}), y) = \mathbb{1}_{[f(\mathbf{x}) \neq y]}$$

Where $\mathbb{1}_{[\pi]}$ is equal to 1 if the predicate π is true and 0 otherwise. In the case of binary classification, the learned function $h : \mathcal{X} \rightarrow \mathbb{R}$ is generally a real-value function and the associated classifier $f : \mathcal{X} \rightarrow \{-1, +1\}$ is defined by taking the signe over the output predictions provided by h . In this case, the equivalent instantaneous loss to the 0/1 loss, defined for the function h is

$$e_0 : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}^+$$

$$(h(\mathbf{x}), y) \mapsto \mathbb{I}_{[y \times h(\mathbf{x}) \leq 0]}$$

Based on the definition of an instantaneous loss and the i. i. d. assumption over the generation of examples with respect to a probability distribution $\mathcal{D}$, the generalization error of a learned function $f \in \mathcal{F}$ is defined as:

$$\mathcal{E}(f) = \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} e(f(\mathbf{x}), y) = \int_{\mathcal{X} \times \mathcal{Y}} e(f(\mathbf{x}), y) d\mathcal{D}(\mathbf{x}, y) \quad (2.1)$$

Where $\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} X(\mathbf{x}, y)$ is the expectation of the random variable X when $(\mathbf{x}, y)$ follows the probability distribution $\mathcal{D}$. As $\mathcal{D}$ is unknown, this generalization error cannot be estimated. In practice, the performance of a function f is estimated over a finite sample set S of size m using its empirical error (or empirical risk) defined as:

$$E(f, S) = \frac{1}{m} \sum_{i=1}^m e(f(\mathbf{x}_i), y_i) \quad (2.2)$$

As examples are generated i. i. d. with respect to $\mathcal{D}$; it is easy to see that the empirical error is an unbiased estimator of the generalization error. Thus, in order to resolve the classification problem for which we have a training set S , it seems natural to choose a function class $\mathcal{F}$ and to search f_S within this class which minimises the empirical error over S .

2.1.2 The Statement of the ERM Principle

This principle is called empirical risk minimization (ERM), and it is the source of many algorithms in machine learning.

The fundamental question which arises here is *according to the ERM principle, is it possible to find a prediction function from a finite set of examples, which has a good generalization performance in all cases?* The answer to this question is, of course, no. For proof, consider the following toy problem

Example **Overfitting**

Suppose that the input dimension is $d = 1$, let the input space $\mathcal{X}$ be the interval $[a, b] \subset \mathbb{R}$ where a and b are real values such that $a < b$, and suppose that the output space is $\{-1, +1\}$. Moreover, suppose that the distribution $\mathcal{D}$ generating the examples $(\mathbf{x}, y)$ is an uniform distribution over $[a, b] \times \{-1\}$. In other terms, examples are randomly chosen over the interval $[a, b]$, and for each observation $\mathbf{x}$ its desired output is -1 .

Consider now, a learning algorithm which minimizes the empirical risk by choosing a function in the function class $\mathcal{F} = \{f : [a, b] \rightarrow \{-1, +1\}\}$ (also denoted as $\mathcal{F} = \{-1, +1\}^{[a, b]}$) in the following way ; after reviewing a training

set $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ the algorithm outputs the prediction function f_S such that

$$f_S(\mathbf{x}) = \begin{cases} -1, & \text{if } \mathbf{x} \in \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \\ +1, & \text{otherwise} \end{cases}$$

In this case, the found classifier has an empirical risk equal to 0, and that for any given training set. However, as the classifier makes an error over the entire infinite set $[a, b]$ except on a finite training set (of zero measure), its generalization error is always equal to 1.

2.2 Consistency of the ERM Principle

The underlying question to the previous question is: *in which case the ERM principle is likely to generate a general learning rule?* The answer of this question lies in a statistical notion called consistency. This concept indicates two conditions that a learning algorithm has to fulfil, namely (a) the algorithm must return a prediction function whose empirical error reflects its generalization error when the size of the training set tends to infinity. Moreover, (b) in the asymptotic case, the algorithm must allow to find the function which minimises the generalization error in the considered function class. Formally:

(a) $\forall \epsilon > 0, \lim_{m \rightarrow \infty} \mathbb{P}(|E(f_S, S) - \mathcal{E}(f_S)| > \epsilon) = 0$, denoted as, $E(f_S, S) \xrightarrow{\mathbb{P}} \mathcal{E}(f_S)$;

(b) $E(f_S, S) \xrightarrow{\mathbb{P}} \inf_{g \in \mathcal{F}} \mathcal{E}(g)$.

These two conditions imply that the empirical error $E(f_S, S)$ of the prediction function found by the learning algorithm over a training S , f_S , converges in probability to its generalization error $\mathcal{E}(f_S)$ and $\inf_{g \in \mathcal{F}} \mathcal{E}(g)$ (Fig. 2.1).

A natural way to analyze the condition (a) of the consistency, expressing the concept of generalization, is to use the following inequality:

$$|\mathcal{E}(f_S) - E(f_S, S)| \leq \sup_{g \in \mathcal{F}} |\mathcal{E}(g) - E(g, S)| \quad (2.3)$$

We see from this inequality that a sufficient condition for generalization is that the empirical risk of the prediction function, whose absolute difference between its empirical and generalization errors among all the functions in $\mathcal{F}$ is the largest, tends to the generalization error of the function, that is:

$$\sup_{g \in \mathcal{F}} |\mathcal{E}(g) - E(g, S)| \xrightarrow{\mathbb{P}} 0 \quad (2.4)$$

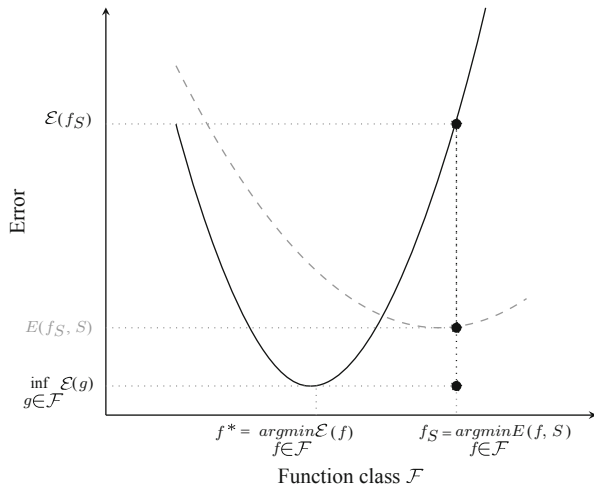


Fig. 2.1 Schematic description of the concept of consistency. The x-axis represents a function class $\mathcal{F}$, the empirical error (dotted line) and the generalization error (solid line) are shown as a function of $f \in \mathcal{F}$. The ERM principle consists in finding a function f_S in the function class $\mathcal{F}$ by minimizing the empirical error over a training set S . The principle is consistent, if $E(f_S, S)$ converges in probability to $E(f_S)$ and $\inf_{g \in \mathcal{F}} E(g)$

This sufficient condition for generalization is a consideration in worse case, and following Eq. 2.3, it implies a bilateral uniform convergence for all functions in the class $\mathcal{F}$. Moreover, the condition (2.4) does not depend to the considered algorithm but just to the function class $\mathcal{F}$. Thus, a necessary condition for the consistency of the ERM principle is that the function class should be restricted (see the overfitting example of the previous section).

The fundamental result of the learning theory (Vapnik 1999, theorem 2.1, p. 38) concerning the consistency of the ERM principle, exhibits another relation involving the supremum over the function class in the form of an unilateral uniform convergence and which stipulates that:

The ERM principle is consistent if and only if:

$$\forall \epsilon > 0, \lim_{m \rightarrow \infty} \mathbb{P} \left(\sup_{f \in \mathcal{F}} [\mathcal{E}(f) - E(f, S)] > \epsilon \right) = 0 \quad (2.5)$$

A direct implication of this result is a uniform bound over the generalization error of all prediction functions $f \in \mathcal{F}$ learned on a training set S of size m and which writes:

$$\forall \delta \in]0, 1], \mathbb{P} (\forall f \in \mathcal{F}, (\mathcal{E}(f) - E(f, S)) \leq \mathfrak{C}(\mathcal{F}, m, \delta)) \geq 1 - \delta \quad (2.6)$$

Where $\mathfrak{C}$ depends on the size of the function class, the size of the training set, and the desired precision $\delta \in]0, 1]$. There are different ways to measure the size

of a function class and the measure commonly used is called complexity or the capacity of the function class. In this chapter, we present two such measures, namely the VC dimension and the Rademacher complexity) leading to different types of generalization bounds and also to the second principle of machine learning which is known as structural risk minimization (SRM).

Before presenting a generalization error bound estimated over a training set which is used to find the prediction function, we will consider in the first instance the estimation of the generalization error of a learned function over a test set (Langford 2005). The aim here is to show that it is possible to have an accurate upper-bound of the generalization error by using a test set that was not used to find the prediction function, and that independently of the capacity of the function class in use.

2.2.1 Estimation of the Generalization Error Over a Test Set

Remind that the examples of a test set are generated i. i. d. with respect to the same probability distribution $\mathcal{D}$ which has generated the training set. Consider f_S a learned function over the training set S , and let $T = \{(\mathbf{x}_i, y_i); i \in \{1, \dots, n\}\}$ be a test set of size n . As the examples of this set are not seen in the training phase, the function f_S does not depend on the values of the instantaneous errors of examples $(\mathbf{x}_i, y_i)$ of this set, and the random variables $(f_S(\mathbf{x}_i), y_i) \mapsto e(f_S(\mathbf{x}_i), y_i)$ can be considered as the independent copies of the same random variable:

$$\begin{aligned} \mathbb{E}_{T \sim \mathcal{D}^n} E(f_S, T) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{T \sim \mathcal{D}^n} e(f_S(\mathbf{x}_i), y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} e(f_S(\mathbf{x}), y) = \mathcal{E}(f_S) \end{aligned}$$

Thus, the empirical error of f_S on the test set, $E(f_S, T)$ is an unbiased estimator of its generalization error.

Furthermore, for each example $(\mathbf{x}_i, y_i)$ let X_i be the random variable $\frac{1}{n} e(f_S(\mathbf{x}_i), y_i)$, as all the random variables $X_i, i \in \{1, \dots, n\}$ are independent and that they take values in $[0, 1]$; by noting that $E(f_S, T) = \sum_{i=1}^n X_i$ and $\mathcal{E}(f_S) = \mathbb{E}\left(\sum_{i=1}^n X_i\right)$, we have the following using Hoeffding (1963) inequality

$$\forall \epsilon > 0, \mathbb{P}([\mathcal{E}(f_S) - E(f_S, T)] > \epsilon) \leq e^{-2n\epsilon^2} \quad (2.7)$$

To better understand this result, let solve the equation $e^{-2n\epsilon^2} = \delta$ with respect to ϵ , hence we have $\epsilon = \sqrt{\frac{\ln 1/\delta}{2n}}$ and:

$$\forall \delta \in]0, 1], \mathbb{P}\left(\mathcal{E}(f_S) \leq E(f_S, T) + \sqrt{\frac{\ln 1/\delta}{2n}}\right) \geq 1 - \delta \quad (2.8)$$

For a small δ , according to Eq. 2.8 we have the following inequality which stands with high probability and all test sets of size n :

$$\mathcal{E}(f_S) \leq E(f_S, T) + \sqrt{\frac{\ln 1/\delta}{2n}}$$

From this result, we have a bound over the generalization error of a learned function which can be estimated using any test set, and in the case where n is sufficiently large, this bound gives a very accurate estimated of the latter.

Example Estimation of the generalization error over a test set (Langford 2005)

Suppose that the empirical error of a prediction function f_S over a test set T of size $n = 1000$ is $E(f_S, T) = 0.23$. For $\delta = 0.01$, i.e. $\sqrt{\frac{\ln(1/\delta)}{2n}} \approx 0.047$, the generalization error of f_S is upperbounded by 0.277 with a probability at least 0.99.

2.2.2 A Uniform Generalization Error Bound

For a given prediction function, we know how to bound its generalization error by using a test set that did not serve to find the parameters of the latter. As part of the study of the consistency of the ERM principle, we would now establish a uniform bound on the generalization error of a learned function depending on its empirical error over a training set. We cannot reach this result, by using the same development than the one presented in the previous section.

This is mainly due to the fact that when the learned function f_S has knowledge of the training data $S = \{(\mathbf{x}_i, y_i); i \in \{1, \dots, m\}\}$, random variables $X_i = \frac{1}{m} e(f_S(\mathbf{x}_i), y_i); i \in \{1, \dots, m\}$ involved in the estimation of the empirical error of f_S on S , are all dependent to each other. Indeed, if we change an example of the training set, the selected function f_S will also change, as well as the instantaneous errors of all the other examples. Therefore, as the random variables X_i can not be considered independently distributed, we are no longer able to apply Hoeffding (1963) inequality.

In the following, we will outline a uniform bound on the generalization error by following the framework of Vapnik (1999). In the next section, we present another, more recent framework for establishing such bounds, developed in the early 2000s by showing the link with the work of Vapnik (1999). For the uniform bound, our starting point is the upper bounding of

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}} [\mathcal{E}(f) - E(f, S)] > \epsilon \right)$$

At this point, there are two cases to consider, the case of finite and infinite sets of functions.

2.2.2.1 Case of Finite Sets of Functions

Consider a function set $\mathcal{F} = \{f_1, \dots, f_p\}$ of size $p = |\mathcal{F}|$. The generalization bound consists in estimating the probability of $\max_{j \in \{1, \dots, p\}} [\mathcal{E}(f_j) - E(f_j, S)]$ being higher than a fixed $\epsilon > 0$, for a given training set of size m .

If $p = 1$, the only choice of selecting the prediction function within $\mathcal{F} = \{f_1\}$ is to take f_1 and that before looking at the examples of any training set S of size m . In this case, we can directly apply the previous bound (Eq. 2.7) from Hoeffding (1963) inequality:

$$\forall \epsilon > 0, \mathbb{P} \left(\max_{j=1} [\mathcal{E}(f_j) - E(f_j, S)] > \epsilon \right) = \mathbb{P} ([\mathcal{E}(f_1) - E(f_1, S)] > \epsilon) \leq e^{-2m\epsilon^2}$$

In the case where $p > 1$, we first note that

$$\max_{j \in \{1, \dots, p\}} [\mathcal{E}(f_j) - E(f_j, S)] > \epsilon \Leftrightarrow \exists f \in \mathcal{F}, \mathcal{E}(f) - E(f, S) > \epsilon \quad (2.9)$$

For a fixed $\epsilon > 0$ and for each function $f_j \in \mathcal{F}$; consider the set of samples of length m on which the generalization error of f_j is larger than its empirical error of more than ϵ :

$$\mathfrak{S}_j^\epsilon = \{S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} : \mathcal{E}(f_j) - E(f_j, S) > \epsilon\}$$

Given a $j \in \{1, \dots, p\}$ and from the previous interpretation, the probability on the samples S of $\mathcal{E}(f_j) - E(f_j, S) > \epsilon$, is less than $e^{-2m\epsilon^2}$, that is

$$j \in \{1, \dots, p\}; \mathbb{P}(\mathfrak{S}_j^\epsilon) \leq e^{-2m\epsilon^2} \quad (2.10)$$

According to the equivalence (2.9), we also have:

$$\begin{aligned} \forall \epsilon > 0, \mathbb{P} \left(\max_{j \in \{1, \dots, p\}} [\mathcal{E}(f_j) - E(f_j, S)] > \epsilon \right) &= \mathbb{P} (\exists f \in \mathcal{F}, \mathcal{E}(f) - E(f, S) > \epsilon) \\ &= \mathbb{P} (\mathfrak{S}_1^\epsilon \cup \dots \cup \mathfrak{S}_p^\epsilon) \end{aligned}$$

We now bound the previous equality by using the result of (2.10), and the union bound:

$$\begin{aligned} \forall \epsilon > 0, \mathbb{P} \left(\max_{j \in \{1, \dots, p\}} [\mathcal{E}(f_j) - E(f_j, S)] > \epsilon \right) \\ = \mathbb{P} (\mathfrak{S}_1^\epsilon \cup \dots \cup \mathfrak{S}_p^\epsilon) &\leq \sum_{j=1}^p \mathbb{P}(\mathfrak{S}_j^\epsilon) \leq p e^{-2m\epsilon^2} \end{aligned}$$

Solving this bound for $\delta = pe^{-2m\epsilon^2}$, i.e. $\epsilon = \sqrt{\frac{\ln(p/\delta)}{2m}}$, and by considering the opposite event we finally obtain:

$$\forall \delta \in]0, 1], \mathbb{P} \left(\forall f \in \mathcal{F}, \mathcal{E}(f) \leq E(f, S) + \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2m}} \right) \geq 1 - \delta \quad (2.11)$$

Compared to the generalization bound obtained on a test set (Eq. 2.8), we see that the empirical error on a test set is a better estimator of the generalization bound; than the empirical error on a training set. In addition, the larger the size of a function class; the higher is the chance that the empirical error on the training set be a significant under-estimate of the generalization error.

Indeed, the interpretation of the bound (2.11), is that for a given $\delta \in]0, 1]$ and for a fraction of possible training sets larger than $1 - \delta$; all the functions of the finite set $\mathcal{F}$ (including the function minimising the empirical error) have a generalization error less than their empirical error plus the residual term $\sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2m}}$. Furthermore, the difference in the worse case between the generalization and the empirical errors tends to 0 when the number of examples tends to infinity and this without making any particular assumption over the distribution of examples $\mathcal{D}$. Thus for any finite set of functions, the ERM principle is consistent for any probability distribution $\mathcal{D}$.

In Sum

The two steps that led to the generalization bound presented in the previous development are as follows:

1. For all fixed function $f_j \in \{f_1, \dots, f_p\}$ and a given $\epsilon > 0$, bound the probability of $\mathcal{E}(f_j) - E(f_j, S) > \epsilon$, over the samples S ;
2. Use the union bound to pass from the probability that holds for each single function in $\mathcal{F}$, to the probability that holds for all the functions in $\mathcal{F}$, at the same time.

2.2.2.2 Case of Infinite Sets of Functions

For the case of an infinite class of functions, the above approach is not directly applicable. Indeed; given a training set of m examples $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$, if we consider the following set:

$$\mathfrak{F}(\mathcal{F}, S) = \{((\mathbf{x}_1, f(\mathbf{x}_1)), \dots, (\mathbf{x}_m, f(\mathbf{x}_m))) \mid f \in \mathcal{F}\} \quad (2.12)$$

The size of this set corresponds to the number of possible ways that the functions of $\mathcal{F}$ can assign class labels to examples $(\mathbf{x}_1, \dots, \mathbf{x}_m)$, and as these functions have two possible outputs (-1 or $+1$), the size of $\mathfrak{F}(\mathcal{F}, S)$ is finite, bounded by 2^m , and this for every considered class function $\mathcal{F}$. Thus a learning algorithm minimizing the

empirical risk on a training set S , chooses the function among $|\mathfrak{F}(\mathcal{F}, S)|$ functions of $\mathcal{F}$ which performs class labels of S resulting in the smallest error. Thereby, there is only a finite number of functions that are involved into the estimation of the empirical error appearing in the expression of

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}} [\mathcal{E}(f) - E(f, S)] > \epsilon \right) \quad (2.13)$$

However, for a different training set S' , the set $\mathfrak{F}(\mathcal{F}, S')$ will be different from $\mathfrak{F}(\mathcal{F}, S)$, and it is impossible to apply the bound for these sets by considering $|\mathfrak{F}(\mathcal{F}, S)|$.

The solution proposed by Vapnik and Chervonenkis is an elegant way to solve this problem, and it consists in replacing the generalization error $\mathcal{E}(f)$ in the expression (2.13) by the empirical error of f on another sample of the same size as S , and called virtual or ghost sample. Formally

Lemma 2.1 (Vapnik and Chervonenkis symmetrization lemma Vapnik (1999))

Let $\mathcal{F}$ be a function class (possibly infinite), and S and S' two training samples of the same size m . For every $\epsilon > 0$, such that $m\epsilon^2 \leq 2$ we have:

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}} [\mathcal{E}(f) - E(f, S)] > \epsilon \right) \leq 2\mathbb{P} \left(\sup_{f \in \mathcal{F}} [E(f, S') - E(f, S)] > \epsilon/2 \right) \quad (2.14)$$

Proof of the symmetrization lemma

Let $\epsilon > 0$ and $f_S^* \in \mathfrak{F}(\mathcal{F}, S)$ the function that realises the supremum $\sup_{f \in \mathcal{F}} [\mathcal{E}(f) - E(f, S)]$. According to the remark above, f_S^* depends on the sample S . We have

$$\begin{aligned} \mathbb{1}_{[\mathcal{E}(f_S^*) - E(f_S^*, S) > \epsilon]} \mathbb{1}_{[\mathcal{E}(f_S^*) - E(f_S^*, S') < \epsilon/2]} &= \mathbb{1}_{[\mathcal{E}(f_S^*) - E(f_S^*, S) > \epsilon \wedge E(f_S^*, S') - \mathcal{E}(f_S^*) \geq -\epsilon/2]} \\ &\leq \mathbb{1}_{[E(f_S^*, S') - E(f_S^*, S) > \epsilon/2]} \end{aligned}$$

By taking the expectation over S' on the previous inequality we have

$$\mathbb{1}_{[\mathcal{E}(f_S^*) - E(f_S^*, S) > \epsilon]} \mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{1}_{[\mathcal{E}(f_S^*) - E(f_S^*, S') < \epsilon/2]} \leq \mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{1}_{[E(f_S^*, S') - E(f_S^*, S) > \epsilon/2]}$$

That is

$$\mathbb{1}_{[\mathcal{E}(f_S^*) - E(f_S^*, S) > \epsilon]} \mathbb{P}(\mathcal{E}(f_S^*) - E(f_S^*, S') < \epsilon/2) \leq \mathbb{E}_{S' \sim \mathcal{D}^m} \mathbb{1}_{[E(f_S^*, S') - E(f_S^*, S) > \epsilon/2]}$$

For each example $(\mathbf{x}'_i, y'_i) \in S'$ denote by X_i the random variable $\frac{1}{m} e(f_S^*(\mathbf{x}'_i), y'_i)$, as f_S^* is independent to the sample S' , the random variables $X_i, i \in \{1, \dots, m\}$ are all independent. The variance of the random variable

Learning with Partially Labeled and Interdependent
Data

Amini, M.-R.; Usunier, N.

2015, XIII, 106 p. 12 illus., Hardcover

ISBN: 978-3-319-15725-2