Bulletin of PFUR Series *Mathematics. Information Sciences. Physics.* No 3. 2008. Pp. 40–54

UDC 519.2, 004.93

Foundational Aspects of Theory of Statistical Function Estimation and Pattern Recognition

E. Fokoué

Department of Mathematics
Kettering University
1700 West Third Avenue, Flint, MI, USA, 48504

This paper provides a gentle introduction to the foundational ideas, concepts and results in the field of science dedicated to the theory of statistical function estimation and pattern recognition. The so-called VC Theory of Vapnik and Chervonenkis is introduced and explored gradually. The emphasis is placed on helping the reader appreciate the importance of the extension of the classical law of large numbers to function spaces, and the key role that "new" concepts such as Empirical Risk Minimization (ERM) principle, ERM consistency, VC-dimension, and complexity control play in constructing algorithms that yield function estimators with optimal properties. As much as possible, each key concept is introduced via a tangible example, with the hope of helping the reader grasp the essential core of the foundational concept under exploration.

Key words and phrases: Statistical Learning Theory, Law of Large Numbers, Consistency, VC Theory, Regularization, Complexity Control, Bounds on generalization, Generalization.

1. Introduction

Let $\mathcal{X}$ and $\mathcal{Y}$ be two sets, and consider their Cartesian product $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$. Now, define $\mathcal{Z}^n \equiv \mathcal{Z} \times \mathcal{Z} \times \ldots \times \mathcal{Z}$ to be the n-fold cartesian product of $\mathcal{Z}$. Assume that $\mathcal{Z}$ is equipped with a probability measure ψ, and let

$$\mathbf{z} \in \mathcal{Z}^n \quad \text{with} \quad \mathbf{z} = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n))$$

denote the realization of a random sample of n examples, where each example $z_i = (\mathbf{x}_i, y_i)$ is independently drawn according to the above probability measure ψ on the product space $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$. Throughout this paper, we consider the following problem: *Given a random sample* $\mathbf{z} = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n))$ *and assuming that the probability measure* ψ *is unknown, find the function* $f : \mathcal{X} \to \mathcal{Y}$ *that best captures the dependencies between the* $\mathbf{x}_i$*'s and the* y_i*'s.* We shall refer to $\mathcal{X}$ as the input space, and to $\mathcal{Y}$ as the output space. To help clarify the key concepts, ideas and results of interest, we shall consider a special case: $\mathcal{X} \subseteq \mathbb{R}^2$ and $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$, corresponding to binary classification in the plane (two dimensional space). Despite its relatively straightforward nature, this special case will provide most of the ingredients needed to address the key foundational issues underlying the theory of statistical function estimation and pattern recognition. Thanks to the fact the binary classification problem is studied using rather basic mathematical tools, the intuition underlying the theory should be followed without too much effort, our emphasis here having been placed on the clarity of the results rather than the technical details thereof. The rest of this paper is organized as follows: in the remainder of section 2, the main definitions and concepts of statistical function estimation are provided. Section 3 explores the concepts introduced in section 2 for the specific case of binary classification in the two dimensional Euclidean space. Section 4 presents the foundational theorems with as much intuitive guidance and examples as possible. Section 4 also touches on some advanced concepts of statistical function estimation such as growth function and VC-dimension, while section 5 is dedicated to the conclusion and the discussion.

2. Loss Functions, Risk Functionals and Optimal Prediction

Def 1 (Loss function and risk functional)**.** Let f denote any generic function mapping an element $\mathbf{x}$ of $\mathcal{X}$ to its corresponding image $f(\mathbf{x})$ in $\mathcal{Y}$. Each time $\mathbf{x}$ is drawn from $\psi(\mathbf{x})$, the disagreement between the image $f(\mathbf{x})$ and the true image y is called the loss, denoted by $L(y, f(\mathbf{x}))$. The expected value of this loss function with respect to the distribution $\psi(\mathbf{x}, y)$ is called the risk functional of f. We shall denote the risk functional of f by $R(f)$, so that

$$R(f) = \mathbb{E}[L(Y, f(X))] = \int L(y, f(\mathbf{x}))\mathrm{d}\psi(\mathbf{x}, y).$$

Best predictor (universal): The best function f^* over the space $\mathcal{Y}^{\mathcal{X}}$ of all measurable functions from $\mathcal{X}$ to $\mathcal{Y}$ is therefore

$$f^* = \mathsf{arg}\inf_f R(f).$$

The risk R^* corresponding to f^* is then defined as

$$R^* = R(f^*) = \inf_f R(f).$$

Best predictor in function class $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$: It turns out in practice that, because of the fact that ψ is unknown, it is hard (almost impossible) to obtain an expression for f^*. One therefore needs to select a function space $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$, and then choose the best estimator $\tilde{f}$ from $\mathcal{F}$, i.e.,

$$\tilde{f} = \mathsf{arg}\inf_{f\in\mathcal{F}} R(f).$$

The risk $\tilde{R}$ associated with the best function in class $\mathcal{F}$ is then defined as

$$\tilde{R} = R(\tilde{f}) = \inf_{f\in\mathcal{F}} R(f).$$

Empirical Risk Minimization Principle: Since the distribution $\psi(\mathbf{x}, y)$ that generates the observations is unknown in practice, the risk functional $R(f)$ which is our criterion for choosing the "best" function $\tilde{f}(\mathbf{x})$ from the function class $\mathcal{F}$ cannot be computed. The theoretical risk functional $R(f)$ is replaced by the so-called empirical risk functional

$$R_n^{\mathbf{z}}(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) \tag{1}$$

based on the random sample $\mathbf{z} = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$. The Empirical Risk Minimization (ERM) principle then consists of finding

$$f_n^{\mathbf{z}} = \mathsf{arg}\min_{f\in\mathcal{F}} R_n^{\mathbf{z}}(f).$$

The slightly complex notation $f_n^{\mathbf{z}}$ is used to emphasize the fact that the estimator obtained via the empirical risk is a random function based on a sample $\mathbf{z}$ of size n. The index n in this case helps to create the sequence of functions when it comes to studying properties such as convergence and rates of convergence. Indeed, a natural question that arises upon realizing that one is dealing with three different types of functions namely f^*, $\tilde{f}$ and $f_n^{\mathbf{z}}$ is: *What is the relationship between these three functions?* For one thing, is it possible to quantify the difference between f^* and $\tilde{f}$, i.e., for some

norm $\|\cdot\|$, what is the value of $\|\tilde{f} - f^*\|$? Since $f_n^{\mathbf{z}}$ is considered as an estimator of $\tilde{f}$, the natural statistical question is then: Is $f_n^{\mathbf{z}}$ a consistent estimator of $\tilde{f}$? And if so, what is the rate of convergence of $f_n^{\mathbf{z}}$ to $\tilde{f}$? One may even be more ambitious and ask instead: Is $f_n^{\mathbf{z}}$ a consistent estimator of f^*? And if so, what is the rate of convergence of $f_n^{\mathbf{z}}$ to f^*? In other words, what can be said about

$$\lim_{n\to\infty} \operatorname*{Prob}_{\mathbf{z}\in\mathcal{Z}^n}\Big\{\|f_n^{\mathbf{z}} - \tilde{f}\| < \varepsilon\Big\}$$

or for that matter

$$\lim_{n\to\infty} \operatorname*{Prob}_{\mathbf{z}\in\mathcal{Z}^n}\Big\{\|f_n^{\mathbf{z}} - f^*\| < \varepsilon\Big\}?$$

It turns out that addressing the comparison between $f_n^{\mathbf{z}}$ and $\tilde{f}$ or $\tilde{f}$ is hard, partly because finding the appropriate norm is not easy, but also constructing bounds is not clear even if one finds such a norm. Fortunately, since all the functions are derived through the risk functional, a more manageable approach to comparing the functions is to compare their corresponding risk functionals. For instance, given a fixed function f, how does $R(f)$ compare to $R_n^{\mathbf{z}}(f)$? Or more formally, for a fixed function f, and for all $\varepsilon > 0$, what is the value of

$$\lim_{n\to\infty} \operatorname*{Prob}_{\mathbf{z}\in\mathcal{Z}^n}\Big\{|R_n^{\mathbf{z}}(f) - R(f)| < \varepsilon\Big\}.$$

In other words, for a fixed given function f, does the empirical risk $R_n^{\mathbf{z}}(f)$ converge to the theoretical risk $R(f)$? This convergence of risk functionals for a fixed function f will be referred to later as *point wise convergence*. Unlike the direct comparison of the functions that ran into problems as mentioned earlier, this comparison via risk functionals provides the added advantage that one can then make confidence statements about the unknown value of $R(f)$. For instance, for a given $0 < \delta < 1$, one can make statements of the form

$$\big|R_n^{\mathbf{z}}(f) - R(f)\big| \leqslant \phi(n, \delta)$$

with a probability of at least $1 - \delta$. Indeed, it turns out that pointwise convergence of the empirical risk to the true risk for a fixed function can be established by rather straightforward application of Chebyshev's inequality.

Theorem 1 (Chebyshev's inequality)**.** *Let ξ be a random variable with finite mean $\mathbb{E}[\xi]$ and finite variance $\sigma^2(\xi) = \mathbb{V}(\xi)$. Then, $\forall \varepsilon > 0$,*

$$\operatorname{Prob}\big\{|\xi - \mathbb{E}(\xi)| > \varepsilon\big\} \leqslant \frac{\mathbb{V}(\xi)}{\varepsilon^2}.$$

A very natural application of Chebyshev's inequality is its use in the study of sums of independent random variables. Indeed, if ξ is a random variable on a probability space $\mathcal{Z}$ with finite mean $\mu = \mathbb{E}[\xi]$ and finite variance $\sigma^2(\xi) = \mathbb{V}(\xi) = \sigma^2$, then $\forall \varepsilon > 0$,

$$\operatorname*{Prob}_{\mathbf{z}\in\mathcal{Z}^n}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) - \mu\right| \geqslant \varepsilon\right\} \leqslant \frac{\sigma^2}{n\varepsilon^2}.$$

An immediate (direct) consequence of Chebyshev's inequality is the fact that the *empirical mean* $\frac{1}{n}\sum_{i=1}^{n}\xi(z_i)$ converges in probability to the *theoretical mean* μ in the limit of very large samples (as $n \to \infty$), i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) \xrightarrow[n\to\infty]{P} \mu.$$

Application to risk functionals: Consider a fixed function $f \in \mathcal{F}$, and let the random variable of interest be ξ with $\xi(z) = L(y_i, f(vx_i))$. Then $R_n^{\mathbf{z}}(f) = (1/n)\sum_{i=1}^n \xi(z_i)$ and $R(f) = \mathbb{E}(\xi)$. Besides, it can be easily shown that $R(f) = \mathbb{E}(R_n^{\mathbf{z}}(f))$. As a result, by Chebyshev's inequality, for a fixed function f, $\forall \varepsilon > 0$,

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{|R_n^{\mathbf{z}}(f) - R(f)| \geqslant \varepsilon\right\} \leqslant \frac{\sigma^2}{n\varepsilon^2},$$

so that, for a fixed function $f \in \mathcal{F}$,

$$R_n^{\mathbf{z}}(f) \xrightarrow[n\to\infty]{P} R(f).$$

Another, and perhaps even more important application of Chebyshev's inequality will be its use in deriving confidence statements about the difference between the empirical quantities of interest and their theoretical counterpart. In other words, while it is important that the convergence occurs, it will be crucial in learning theory to know how fast the convergence is in terms of (a) the number of examples observed (sample size); (b) the confidence level desired $(1-\delta)$; and (c) a characteristic of the function class under consideration, maybe through its quantities like variances and other moments. For instance, the above Chebyshev's inequality on sums of random variables can be rewritten as

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\left|\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mu\right| \leqslant \sqrt{\frac{\sigma^2}{n\delta}}\right\} \leqslant 1-\delta,$$

by simply setting $\varepsilon = \sqrt{\sigma^2/(m\delta)}$ for any $0 < \delta < 1$. Therefore, one could assert, that with confidence at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mu\right| < \sqrt{\frac{1}{\delta}\frac{\sigma^2}{n}}.$$

Bounds and rates for fixed f: Applying the above to risks functionals yields, for a fixed $f \in \mathcal{F}$,

$$|R_n^{\mathbf{z}}(f) - R(f)| < \sqrt{\frac{1}{\delta}\frac{\sigma^2}{n}}.$$

As it turns out, extensions (improvements) on Chebyshev's inequality will yield faster rates of convergence of empirical quantities $R_n^{\mathbf{z}}(f)$ to their theoretical counterparts $R(f)$. One such improvement is provided by Hoeffding's inequality.

Theorem 2 (Hoeffding's inequality)**.** *Let $\xi(z_1), \xi(z_2), \ldots, \xi(z_n)$ be a collection of i.i.d random variables with $\xi(z_i) \in [a,b]$. Then, $\forall \varepsilon > 0$,*

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi)\right| > \varepsilon\right] \leqslant 2\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right).$$

For all $\delta \in (0,1)$, Hoeffding's inequality allows one to write

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi)\right| > (b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right] \leqslant \delta,$$

so that with probability at least $1-\delta$,

$$\left|\frac{1}{n}\sum_{i=1}^n \xi(z_i) - \mathbb{E}(\xi)\right| \leqslant (b-a)\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

Assuming the loss function is bounded, a direct application of Hoeffding's inequality to risks functionals for a fixed $f \in \mathcal{F}$ yields,

$$|R_n^{\mathbf{z}}(f) - R(f)| < (b-a)\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

By rewriting the Chebyshev bound as

$$|R_n^{\mathbf{z}}(f) - R(f)| < \sqrt{\frac{2}{\delta}\frac{1}{2n}\sigma^2}. \tag{2}$$

it appears clear that Hoeffding provides faster rates of convergence, since the term in δ is changed from $2/\delta$ to $\ln(2/\delta)$, and also instead of the variance σ^2 of ξ, we now $(b-a)$. Most of the that bounds are encountered here will have in them a function of $2/\delta$, a function of $1/2n$ and a function of the variance. Indeed, $1-\delta$ confidence bounds will be typically of the form

$$c\left[\frac{\ln(2/\delta)}{n}\right]^{\frac{1}{2}+\theta}$$

with $0 < \theta < \frac{1}{2}$ depending on the variance of the random variable ξ.

For all the good insights they help gain, the above bounds, by virtue of their point wise nature, suffer from the following limitations: Given two distinct functions f and g both from $\mathcal{F}$, the set $\mathbf{Z}_f$ of samples for which the probabilistic inequality holds for fixed f may differ from the $\mathbf{Z}_g$ of samples for which the probabilistic inequality holds for fixed g. In other words, crucially needed comparisons of the type

$$|R_n^{\mathbf{z}}(f) - R(g)| \quad \text{or} \quad |R_n^{\mathbf{z}}(g) - R(f)|$$

are not taken into account in the study of pointwise convergence. Clearly, more important than point wise convergence is *uniform convergence* that helps compare the risk functionals, not just for a fixed function, but for functions across the space of functions under consideration. For instance, it is crucial to quantify or at least provide a bound for the difference between the empirical risk $R_n^{\mathbf{z}}(f_n^{\mathbf{z}})$ and the smallest risk $\tilde{R} = R(\tilde{f})$ in the function class $\mathcal{F}$ under consideration, but also the difference between $R_n^{\mathbf{z}}(f_n^{\mathbf{z}})$ and $\tilde{R} = R(\tilde{f})$. In other words, one needs to study random sequences like

$$\left|R(f_n^{\mathbf{z}}) - \inf_{f\in\mathcal{F}} R(f)\right| \quad \text{and} \quad \left|R_n^{\mathbf{z}}(f_n^{\mathbf{z}}) - \inf_{f\in\mathcal{F}} R(f)\right|$$

to gain insights into the quality of function estimation provided by $f_n^{\mathbf{z}}$ through the ERM principle. More generally, it boils to investigating various aspects of

$$\lim_{n\to\infty} \operatorname*{Prob}_{\mathbf{z}\in\mathcal{Z}^n}\left\{\sup_{f\in\mathcal{F}}\left|R_n^{\mathbf{z}}(f) - R(f)\right| < \varepsilon\right\}.$$

A reasoning on error decomposition and consistency of estimators along with rates, bounds and algorithms applies to function spaces: indeed, the difference between the true risk $R(f_n^{\mathbf{z}})$ associated with $f_n^{\mathbf{z}}$ and the overall minimum risk R^* can be decomposed to explore in greater details the source of error in the function estimation process:

$$R(f_n^{\mathbf{z}}) - R^* = \underbrace{R(f_n^{\mathbf{z}}) - R(\tilde{f})}_{\text{Estimation error}} + \underbrace{R(\tilde{f}) - R^*}_{\text{Approximation error}}. \tag{3}$$

Theorem 3 (Consistency of the Empirical Risk Minimization principle)**.** *The ERM principle is* consistent *if it provides a sequence of functions* $f_n^{\mathbf{z}}$, $n = 1, 2, \ldots$ *for which both the expected risk* $R(f_n^{\mathbf{z}})$ *and the empirical risk* $R_n^{\mathbf{z}}(f_n^{\mathbf{z}})$ *converge to the minimal possible value of the risk* $R(\tilde{f})$ *in the function class under consideration, i.e.,*

$$R(f_n^{\mathbf{z}}) \xrightarrow[n\to\infty]{P} \inf_{f\in\mathcal{F}} R(f) = R(\tilde{f}) \quad \textit{and} \quad R_n^{\mathbf{z}}(f_n^{\mathbf{z}}) \xrightarrow[n\to\infty]{P} \inf_{f\in\mathcal{F}} R(f) = R(\tilde{f}).$$

It turns out the above theorem on the consistency of the empirical risk minimization principle constitutes one of the four pillars of statistical learning theory as formulated and presented by such authors as [1]. When constructing function estimators, the least one should do is assess the convergence of the empirical quantities of interest to their theoretical counterparts. In [1], Vapnik provides the following four questions as the keys to statistical learning theory (a) *What are the necessary and sufficient conditions for the consistency of a learning process based on the ERM principle?* This first questions suggests the need for a *theory of consistency of learning processes.* (b) How fast is the rate of convergence of the learning process? A question that opens the door to the need for a *nonasymptotic theory of the rate of convergence of learning processes* as opposed to the traditional - sometimes unrealistic - asymptotic theory. (c) How can one control the rate of convergence (the generalization ability) of the learning process?. Here, the implication of the question is the need to develop tools and a *theory for controlling the generalization ability of learning processes*; and finally, (d) How can one construct algorithms that can control the generalization ability of the learning process?. This last question, clearly of interest to practitioners, allows statistical learning theory to seek to provide *tools along with a theory for constructing/devising learning algorithms*, with the aim of consolidating all the four pillars. Algorithms constructed this way are expected to focus on the problem at hand, with all its aspects taken into account as thoroughly as possible. In [1], Vapnik discusses the details of this theorem at length, and extends the exploration to include the difference between what he calls trivial consistency and non-trivial consistency. To better understand consistency in function spaces, consider the sequence of random variables

$$\omega^n = \sup_{f\in\mathcal{F}} |R(f) - R_n^{\mathbf{z}}(f)|, \tag{4}$$

and consider studying

$$\lim_{n\to\infty} P\left\{ \sup_{f\in\mathcal{F}} |R(f) - R_n^{\mathbf{z}}(f)| > \varepsilon \right\} = 0, \quad \forall \varepsilon > 0.$$

Vapnik shows that the sequence of the means of the random variable ξ^n converges to zero as the number n of observations increases [1]. He also remarks that the sequence of random variables ξ^n converges in probability to zero if the set of functions $\mathcal{F}$, contains a finite number N of elements. We will show that later in the case of pattern recognition. It remains then to *describe the properties of the set of functions* $\mathcal{F}$*, and probability measure* $\psi(\mathbf{x}, y)$ *under which the sequence of random variables* ξ^n *converges in probability to zero.*

$$\lim_{n\to\infty} P\left\{ \left[\sup_{f\in\mathcal{F}} [R(f) - R_n^{\mathbf{z}}(f)] > \varepsilon \right] \text{ or } \left[\sup_{f\in\mathcal{F}} [R_n^{\mathbf{z}}(f) - R(f)] > \varepsilon \right] \right\} = 0.$$

3. Statistical Theory of Pattern Recognition

Instead of exploring the details of the above theorem in pure abstraction, one of the most commonly encountered statistical tasks will now be explored under the framework of statistical learning theory and the details of the theorem will be clarified along the way: that task is statistical pattern recognition, and in this case binary

classification in two dimensional space will be studied. For this problem, the so-called $0-1$ loss function defined below is used. More specifically,

$$L(y, f(\mathbf{x})) = \mathbf{I}(y \neq f(\mathbf{x})) = \begin{cases} 0 & \text{if} \quad y = f(\mathbf{x}), \\ 1 & \text{if} \quad y \neq f(\mathbf{x}). \end{cases} \tag{5}$$

Now, with the zero-one loss function in binary classification, our corresponding true risk (also known as theoretical risk or generalization error or true error) is given by

$$R(f) = \int L(y, f(\mathbf{x}))d\psi(\mathbf{x}, y) = \mathbb{E}\left[\mathbf{I}(Y \neq f(X))\right] = \text{Prob}_{(X,Y)\sim\psi}[Y \neq f(X)]. \tag{6}$$

The true error $R(f)$ of a classifier f therefore defines the probability that f misclassifies any arbitrary observation randomly draw from the population of interest according to the distribution ψ. It is important to note from the definition that $R(f)$ can also be interpreted as the *expected* disagreement between classifier f and the truth about the label y of $\mathbf{x}$.

How is the Bayes' classifier obtained? Consider a pattern $\mathbf{x}$ from the input space, and a class label y. Let $p(\mathbf{x}|y)$ denote the class conditional density of $\mathbf{x}$ in class y, and let $\text{Prob}[Y = y]$ denote the prior probability of class membership. The posterior probability of class membership is defined as

$$\text{Prob}[Y = y|\mathbf{x}] = \frac{\text{Prob}[Y = y]p(\mathbf{x}|y)}{p(\mathbf{x})}.$$

Given a pattern $\mathbf{x}$ to be classified, the Bayes classification strategy consists of

$$\text{Assigning } \mathbf{x} \text{ to the class with } \mathbf{maximum} \text{ posterior probability.}$$

More formally, with f^* denoting the function from $\mathcal{X}$ to $\{-1, +1\}$ that implements the Bayes classifier, we have,

$$f^*(\mathbf{x}) \quad = \quad \begin{cases} +1, & \text{if } \text{Prob}[Y = +1|\mathbf{x}] = \max\limits_{y\in\{-1,+1\}} \left\{\text{Prob}[Y = y|\mathbf{x}]\right\}, \\ -1, & \text{if } \text{Prob}[Y = -1|\mathbf{x}] = \max\limits_{y\in\{-1,+1\}} \left\{\text{Prob}[Y = y|\mathbf{x}]\right\}. \end{cases}$$

For simplicity, we shall write

$$f^*(\mathbf{x}) \quad = \quad \arg \max_{c\in\{-1,+1\}} \text{Prob}(Y = c|\mathbf{x}). \tag{7}$$

Theorem 4. *The minimizer of the* $0-1$ *risk functional over all possible classifiers is the Bayes classifier* f^* *defined above. Therefore, the Bayes' classifier* f^* *is such that*

$$f^*(\mathbf{x}) = \arg \inf_f \text{Prob}_{(X,Y)\sim\psi}[Y \neq f(X)] = \arg \inf_f \mathbb{E}\left[\mathbf{I}(Y \neq f(X))\right].$$

Proof. Given $\mathbf{x}$, the conditional risk (risk given $\mathbf{x}$) can be broken down as follows: the conditional risk of assigning $\mathbf{x}$ to class $+1$ is

$$\begin{aligned} R(f(\mathbf{x}) = +1) \quad &= \quad L(f(\mathbf{x}) = +1, Y = +1)\text{Prob}[Y = +1|\mathbf{x}] + \\ &+ \quad L(f(\mathbf{x}) = +1, Y = -1)\text{Prob}[Y = -1|\mathbf{x}] = \\ &= \quad \text{Prob}[Y = -1|\mathbf{x}] = 1 - \text{Prob}[Y = +1|\mathbf{x}]. \end{aligned}$$

From the above, minimizing the risk $R(f(\mathbf{x}) = +1)$ of deciding to assign $\mathbf{x}$ to class $+1$ under the 0-1 loss is equivalent to maximizing the posterior probability $\text{Prob}[Y = +1|\mathbf{x}]$ of $\mathbf{x}$ being in class $+1$. Therefore, the function f that minimizes $R(f(\mathbf{x}) = +1)$

is the same function that is based on $\text{Prob}[Y = +1|\mathbf{x}]$ being the maximum. With all that, if

$$g = \arg\min_f R(f(\mathbf{x}) = +1),$$

then

$$g(\mathbf{x}) = \arg\max_{c\in\{-1,+1\}} \text{Prob}(Y = c|\mathbf{x}) = f^*(\mathbf{x}).$$

The same reasoning can be made for $R(f(\mathbf{x}) = -1)$. The Bayes classifier is indeed the universal minimizer of the 0-1 risk functional. □

Note that the definition and therefore the construction of the Bayes classifier requires the knowledge of the probability density $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$ which in practice is unknown. As stated in section 1, one then has to select a space of functions that one assumes contains a good classifier of the data.

A Simple Two Dimensional Classification Task

Consider the classification task of Figure 1 as an illustrative example. In order to construct the Bayes classifier for this task, one needs to know the probability measure according to which the points are generated.

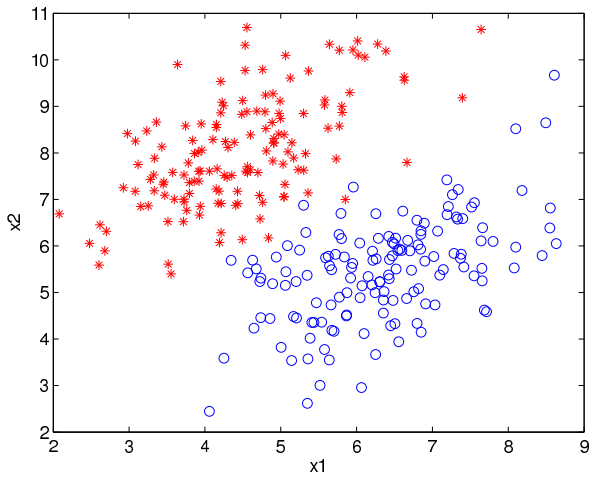

Figure 1. **Binary classification task in a two dimensional space**

One way to circumvent the fact of not knowing $\psi(\mathbf{x}, y)$ is to estimate the density $p(\mathbf{x}, y)$ and then construct the approximate Bayes' classifier. However, density estimation, as warned by Vapnik [1] and many other authors is, not only a hard problem in it own right, but an ill-posed problem also. In fact, a standard wisdom promoted by Vapnik in statistical learning theory is to solve the classification problem as directly as possible and avoid complicating the task with many intermediary and often hard problems. Instead of trying the construct the overall best classifier for the task, one should consider restricting the function search to a class of classifiers. In this case, the scatter seems to suggest that linear separation might be a decent strategy. In other words, one may consider finding the best linear separating hyperplane, i.e.

$$\mathcal{F} = \Big\{ f : \mathcal{X} \to \{-1, +1\} | \exists \alpha_0 \in \mathbb{R}, (\alpha_1, \ldots, \alpha_p)^\top = \boldsymbol{\alpha} \in \mathbb{R}^p | \\ f(\mathbf{x}) = \text{sign}\left(\boldsymbol{\alpha}^\top \mathbf{x} + \alpha_0\right), \forall \mathbf{x} \in \mathcal{X} \Big\}.$$

It is important to note that although the function spaces in the above examples are driven by parameters that are components of a vector, $\boldsymbol{\alpha}$ need not be a vector for more

general function spaces. In fact, $\boldsymbol{\alpha}$ is allowed to be any abstract set of parameters, so that any arbitrary set of functions can be defined and handled by this framework. Now, since the true risk for the best function in this class $\mathcal{F}$ cannot be computed because of the fact that $\psi(\mathbf{x}, y)$ is unknown, one has to turn to the empirical risk. The empirical risk or empirical error corresponding to the true risk of equation (6), is given by

$$R_n^{\mathbf{z}}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(y_i \neq f(\mathbf{x}_i)) = \text{Prob}_{(\mathbf{x},y)\sim \text{S}}\left[Y \neq f(X)\right]. \tag{8}$$

$\mathbf{I}$ is the indicator function, and $\text{Prob}_{(\mathbf{x},y)\sim \text{S}}[\ldots]$ is a probability taken with respect to the uniform distribution over the sample S. $R_n^{\mathbf{z}}(f)$ is therefore the *realized* disagreement between classifier f and the truth about the label y of $\mathbf{x}$ based on information contained in the sample S. In the case of the example of Figure 1, the function

$$f_n^{\mathbf{z}} = \arg\min_{f\in\mathcal{F}} R_n^{\mathbf{z}}(f)$$

is the straight line that separates the two classes with the minimum number of misclassifications. It is crucial to mention that, in the presence of the data of Figure 1, seeking to minimize the empirical risk without restricting the function class does lead overfitting. Indeed, one could construct a classifier that achieves zero empirical risk, but such a classifier, not only is too complex, but also does not generalize well, in the sense that future observations are likely to be misclassified. The control of complexity mentioned earlier as one of the pillars of statistical learning theory does come into play in such cases. Statistical learning theory deals with this issue by way of the so-called *Structural risk minimization principle* [1]. For now, the focus is on the convergence of the ERM principle in pattern recognition. It is easy to see that, for a given (fixed) function (classifier) f,

$$\mathbb{E}[R_n^{\mathbf{z}}(f)] = R(f). \tag{9}$$

Remember that the goal of statistical function estimation is to devise a technique (strategy) that chooses from the function class $\mathcal{F}$, the one function whose true risk is as close as possible to the lowest risk in class $\mathcal{F}$. The question arises: since one cannot calculate the true error, how can one devise a learning strategy for choosing classifiers based on it? Tentative answer: At least devise strategies that yield functions for which the upper bound on the theoretical risk is as tight as possible, so that one can make confidence statements of the form:

With probability $1-\delta$ *over an i.i.d. draw of some sample according to the distribution* ψ*, the expected future error rate of some classifier is bounded by some function* $\phi(\delta, \text{error rate on sample})$ *of* δ *and the error rate on sample.*

If one resort to Chebyshev's inequality encountered earlier, it is easy to see that

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\left|R_n^{\mathbf{z}}(f) - R(f)\right| > \varepsilon\right\} \leqslant \frac{R(f)(1-R(f))}{n\varepsilon^2}$$

for a given classifier f. Since $\max\limits_{R(f)\in[0,1]} R(f)(1-R(f)) = \frac{1}{4}$, we have

$$\sqrt{\frac{R(f)(1-R(f))}{n\delta}} < \sqrt{\frac{1}{4n\delta}}.$$

Based on Chebyshev's inequality, for a given classifier f, with a probability of at least $1-\delta$, the bound on the difference between the true risk $R(f)$ and the empirical risk $R_n^{\mathbf{z}}(f)$ is given by

$$\left|R_n^{\mathbf{z}}(f) - R(f)\right| < \sqrt{\frac{1}{4n\delta}}.$$

A tighter bound is derived from Hoeffding's inequality mentioned earlier. More specifically, for a fixed function f and for any $\delta \in (0,1)$,

$$R(f) \leqslant R_n^{\mathbf{z}}(f) + \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}.$$

with probability at least $1-\delta$.

Fact 1. *The bound yielded by Hoeffding's inequality is tighter than the one derived from Chebyshev's inequality.*

Proof. Clearly, we need to find out which of $\ln 2/\delta$ or $1/2\delta$ is larger. This is the same as comparing $\exp(1/2\delta)$ and $2/\delta$, which in turns means comparing $a^{(2/\delta)}$ and $2/\delta$ where $a = \exp(1/4)$. With $\delta > 0$, $a^{(2/\delta)} > 2/\delta$, so that, we know that Hoeffding's bounds are tighter. The graph also confirms this. □

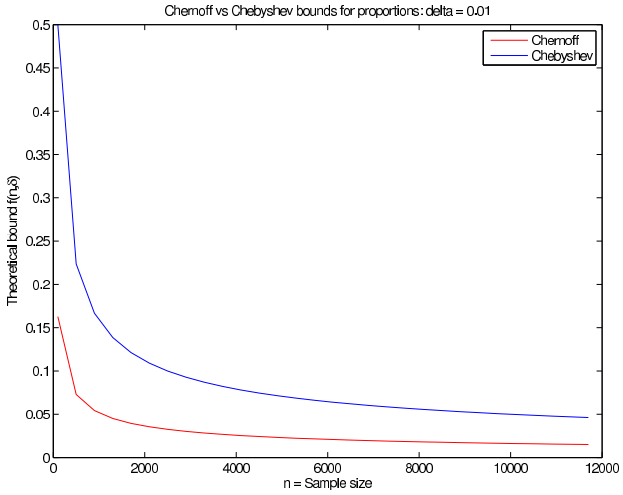

Figure 2. **Chernoff vs Chebyshev bounds for proportions:** $\delta = 0.01$

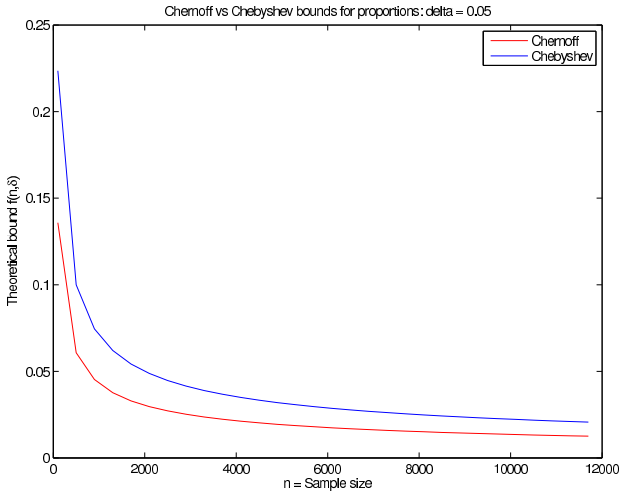

Figure 3. **Chernoff vs Chebyshev bounds for proportions:** $\delta = 0.05$

In all the above, we only addressed pointwise convergence of $R_n^{\mathbf{z}}(f)$ to $R(f)$, i.e., for a fixed machine $f \in \mathcal{F}$, we studied the convergence of

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\Big\{\big|R_n^{\mathbf{z}}(f) - R(f)\big| > \varepsilon\Big\} \quad \text{to} \quad 0.$$

Needless to mention that pointwise convergence is of very little use here for reasons mentioned in great detail in section 1. Indeed, when one writes *for a fixed function f and for any $\delta \in (0,1)$,*

$$R(f) \leqslant R_n^{\mathbf{z}}(f) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

with probability at least $1-\delta$, one actually means the following: If we choose a fixed function f and then collect many different samples $\mathbf{z} \in \mathcal{Z}^n$, then the corresponding empirical risks for most of those samples *a proportion of at least* $1-\delta$ *of them* will be close to the true risk. However, for a fixed sample $\mathbf{z} \in \mathcal{Z}^n$, one can find a function for which the difference between the empirical risk and the theoretical risk is very large, especially if the function class $\mathcal{F}$ is large enough. Therefore, instead, for bounds to be useful, they have to apply to all function in $\mathcal{F}$, not just pointwise. A more interesting issue to address is uniform convergence. That is, for all machines, $f \in \mathcal{F}$, determine the necessary and sufficient conditions for the convergence of

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\sup_{f\in\mathcal{F}}|R_n^{\mathbf{z}}(f) - R(f)| > \varepsilon\right\} \quad \text{to} \quad 0.$$

4. Derivation of Bounds on the Generalization Error

4.1. Bounds for Finite Function Classes

Suppose that the function class $\mathcal{F}$ has m functions in it, i.e., $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$. One seeks to derive bounds that apply at once to all the functions in $\mathcal{F}$. In other words, instead of stopping at the inequality

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\Big\{\big|R_n^{\mathbf{z}}(f) - R(f)\big| > \varepsilon\Big\} \leqslant 2\mathrm{e}^{-2n\varepsilon^2},$$

one needs to compute the proportion

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\sup_{f\in\mathcal{F}}|R_n^{\mathbf{z}}(f) - R(f)| > \varepsilon\right\} = \underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\Big\{\exists f \in \mathcal{F} : \ \big|R_n^{\mathbf{z}}(f) - R(f)\big| > \varepsilon\Big\}.$$

The supremum of all deviations is greater than ε if there exists at least one function whose deviation is greater than ε.

Lemma 1. *The function class $\mathcal{F}$ has m functions in it, i.e., $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$, then Hoeffding's inequality is extended to the supremum so that, $\forall \varepsilon > 0$,*

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\sup_{f\in\mathcal{F}}|R_n^{\mathbf{z}}(f) - R(f)| > \varepsilon\right\} \leqslant 2m\mathrm{e}^{-2n\varepsilon^2}.$$

Proof.

$$\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\sup_{f\in\mathcal{F}}|R_n^{\mathbf{z}}(f) - R(f)| > \varepsilon\right\} = \underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\exists f \in \mathcal{F} : \ \big|R_n^{\mathbf{z}}(f) - R(f)\big| > \varepsilon\right\} =$$

$$= \underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\left\{\bigcup_{i=1}^{m}\big\{\big|R_n^{\mathbf{z}}(f_i) - R(f_i)\big| > \varepsilon\big\}\right\} \leqslant \sum_{i=1}^{m}\underset{\mathbf{z}\in\mathcal{Z}^n}{\text{Prob}}\big\{\big|R_n^{\mathbf{z}}(f_i) - R(f_i)\big| > \varepsilon\big\} \leqslant 2m\mathrm{e}^{-2n\varepsilon^2}.$$

□

Proposition 1. *If the function class $\mathcal{F}$ is finite, i.e. $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$, where $m = |\mathcal{F}| = \#\mathcal{F} =$ Number of functions in the class $\mathcal{F}$, then, for all $f \in \mathcal{F}$,*

$$R(f) \leqslant R_n^{\mathbf{z}}(f) + \left(\frac{\ln m + \ln \frac{2}{\delta}}{2n}\right)^{1/2}$$

with probability at least $1 - \delta$, $\forall \delta > 0$.

Proof. It is obvious that for all $f \in \mathcal{F}$,

$$|R_n^{\mathbf{z}}(f) - R(f)| \leqslant \sup_{f \in \mathcal{F}} |R_n^{\mathbf{z}}(f) - R(f)|.$$

Therefore, by virtue of the above lemma, $\forall f \in \mathcal{F}$,

$$|R_n^{\mathbf{z}}(f) - R(f)| \leqslant 2m\mathrm{e}^{-2n\varepsilon^2}.$$

For all $\delta \in (0, 1)$, setting $\delta = 2m\mathrm{e}^{-2n\varepsilon^2}$ yields

$$\forall f \in \mathcal{F},\ |R_n^{\mathbf{z}}(f) - R(f)| \leqslant \sqrt{\frac{\ln m + \ln \frac{2}{\delta}}{2n}}$$

with probability at least $1 - \delta$, as required. □

Since the above result applies to all the functions in $\mathcal{F}$, it therefore applies to

$$f_n^{\mathbf{z}} = \arg\min_{f \in \mathcal{F}} R_n^{\mathbf{z}}(f)$$

which is the function constructed by the algorithm to classify the data. One can therefore safely write,

$$R(f_n^{\mathbf{z}}) \leqslant R_n^{\mathbf{z}}(f_n^{\mathbf{z}}) + \left(\frac{\ln m + \ln \frac{2}{\delta}}{2n}\right)^{1/2},$$

thereby providing a bound on the true error for the classifier in hand. The term $\ln m$ reflects the fact that the bound on the generalization error must hold for all the functions in the class $\mathcal{F}$.
While the above results help compare the empirical risk to the theoretical risk across the function class, there remains the need to compare the true risk of the constructed function $R(f_n^{\mathbf{z}})$ to the smallest risk in the class or the smallest risk overall. The following theorem helps do just that.

Theorem 5. *If $R_n^{\mathbf{z}}(f)$ and $R(f)$ are close for all $f \in \mathcal{F}$, i.e., $\forall \varepsilon > 0$,*

$$\sup_{f \in \mathcal{F}} |R_n^{\mathbf{z}}(f) - R(f)| \leqslant \varepsilon, \text{ then } R(f_n^{\mathbf{z}}) - R(\tilde{f}) \leqslant 2\varepsilon.$$

Proof. Recall that we did define $f_n^{\mathbf{z}}$ as the best function that is yielded by the empirical risk $R_n^{\mathbf{z}}(f)$ in the function class $\mathcal{F}$. Recall also that $R_n^{\mathbf{z}}(f_n^{\mathbf{z}})$ can be made as small as possible as we saw earlier. Therefore, with $\tilde{f}$ being the best true risk in class $\mathcal{F}$, we always have

$$R_n^{\mathbf{z}}(\tilde{f}) - R_n^{\mathbf{z}}(f_n^{\mathbf{z}}) \geqslant 0.$$

As a result,

$$R(f_n^{\mathbf{z}}) = R(f_n^{\mathbf{z}}) - R(\tilde{f}) + R(\tilde{f}) = R_n^{\mathbf{z}}(\tilde{f}) - R_n^{\mathbf{z}}(f_n^{\mathbf{z}}) + R(f_n^{\mathbf{z}}) - R(\tilde{f}) + R(\tilde{f}) \leqslant$$

$$\leqslant 2\sup_{f\in\mathcal{F}}|R(f)-R_n^{\mathbf{z}}(f)|+R(\tilde{f}).$$

Consequently,

$$R(f_n^{\mathbf{z}})-R(\tilde{f})\leqslant 2\sup_{f\in\mathcal{F}}|R(f)-R_n^{\mathbf{z}}(f)|$$

as required. □

Corollary 1. A direct consequence of the above theorem is the following:

$$R(f_n^{\mathbf{z}})\leqslant R(\tilde{f})+2\left(\frac{\ln m+\ln\frac{2}{\delta}}{2n}\right)^{1/2} \tag{10}$$

with probability at least $1-\delta$, $\forall\delta>0$, where as before

$$\tilde{f}=\arg\inf_{f\in\mathcal{F}}R(f)\quad\text{and}\quad f_n^{\mathbf{z}}=\operatorname{argmin}_{f\in\mathcal{F}}R_n^{\mathbf{z}}(f).$$

Equation (10) is of foundational importance, because it reveals clearly that the size of the function class controls the uniform bound on the crucial generalization error: Indeed, *if the size m of the function class $\mathcal{F}$ increases, then $R(\tilde{f})$ is caused to increase while $R(f_n^{\mathbf{z}})$ decreases*, so that the trade-off between the two is controlled by the size m of the function class. This raises the natural question as to what happens when the function class is infinite dimensional. Indeed, for infinite dimensional function spaces, one will need to introduce such concepts as the capacity of the function space, measured through devices such as the VC-dimension and covering numbers.

4.2. Statistical Theory for Infinite Dimensional Spaces

When the function class $\mathcal{F}$ is uncountable, one way to define its "size" is by way of the samples on which the functions from that class operate. In the case of binary classification for instance, the number of labellings $(-1,+1)$ yielded by separating hyperplanes on a given sample $\mathbf{z}\in\mathcal{Z}^n$ of size n is finite even though the number of hyperplanes itself is infinite. One could define

$$\mathcal{F}_{\mathbf{z}}=\{(f(z_1),f(z_2),\ldots,f(z_n)):f\in\mathcal{F}\}$$

to be the set of ways in which $\mathbf{z}=(z_1,z_2,\ldots,z_n)$ can be classified.

Def 2. The growth function is the maximum number of ways in which n points can be classified by a function class. More specifically,

$$S_{\mathcal{F}}(n)=\sup_{\mathbf{z}\in\mathcal{Z}^n}|\mathcal{F}_{\mathbf{z}}|.$$

For the binary classification task, there are 2^n ways to classify a sample of size n into two classes $\{=1,+1\}$. Therefore, for binary classification $S_{\mathcal{F}}(n)\leqslant 2^n$ for any function class $\mathcal{F}$. The following theorem by Vapnik and Chervonenkis plays a foundational role in statistical learning theory.

Theorem 6 (Vapnik-Chervonenkis)**.** *For any $\delta\in(0,1)$,*

$$\forall f\in\mathcal{F},\quad R(f)\leqslant R_n^{\mathbf{z}}(f)+2\sqrt{2\frac{\ln S_{\mathcal{F}}(2n)+\ln\frac{2}{\delta}}{2n}}$$

with probability at least $1-\delta$.

It is therefore crucial to be able to compute the growth function $S_{\mathcal{F}}(n)$ of a given function class $\mathcal{F}$. The concept of VC-dimension provides a way to manipulate the growth function.

Def 3 (VC-dimension)**.** The VC-dimension h of a function class $\mathcal{F}$ is the largest n such that

$$S_{\mathcal{F}}(n) = 2^n.$$

It is shown [1] that, if $\mathcal{F}$ is the class of separating hyperplanes in a p dimensional space, then

$$VC\dim(\mathcal{F}) = h = p + 1.$$

The VC-dimension and the growth function for that matter can be viewed as as measures of the effective size of a function class. In a sense, by "projecting" the function class onto a finite sample, one avoids simply counting the number of functions in that class, and instead one captures the geometry of the function class and can therefore compute finite quantity that measure the size of that class relative to the finite sample.

The question remains: how does the VC-dimension help provide bounds on the generalization error when the function class is infinite dimensional. The answer requires the following lemma.

Lemma 2 (Vapnik and Chervonenkis, Sauer, Shelah)**.** *Let $\mathcal{F}$ be a function class with finite VC-dimension h. Then*

$$S_{\mathcal{F}}(n) \leqslant \sum_{i=0}^{n} \binom{n}{i}, \text{ and for all } n \geqslant h \quad S_{\mathcal{F}}(n) \leqslant \left(\frac{en}{h}\right)^h.$$

From the above lemma, the following result is derived immediately: If a function class $\mathcal{F}$ has finite VC-dimension h, then for all $\delta \in (0, 1)$,

$$\forall f \in \mathcal{F}, \quad R(f) \leqslant R_n^{\mathbf{z}}(f) + 2\sqrt{2\frac{h\left(\frac{\ln 2n}{h} + 1\right) + \ln\frac{2}{\delta}}{n}} \tag{11}$$

with probability at least $1 - \delta$. As a consequence of the above result, it turns out that the difference between the empirical risk and the true risk of order at most $\sqrt{(h \ln n)/n}$. A finite VC-dimension ensures that the empirical risk converges uniformly over the class to the true risk.

5. Conclusion and Discussion

This paper has provided a general introduction to the theory underlying statistical function estimation and pattern recognition. As mentioned in section 1, there are indeed four pillars of statistical learning theory, of which we have touched on the first two, namely, (a) the necessary and sufficient conditions for the consistency of the Empirical Risk Minimization (ERM) principle, and (a) the derivation of non-asymptotic rates of convergence thereof. The remaining two aspects of the foundation, namely (c) the control of complexity and (d) the construction of learning algorithms requires a substantial amount of space. The introduction provided here sheds enough light onto the usefulness and the challenges of this field. Clearly, while it is good to build classifiers, it is crucial to study their theoretical properties, and that's what this field provides.

The field is vast and the topics and variety and many. A more detailed account of the topic with applications and more advanced theoretical developments can be found in the cited references.

References

1. *Vapnik V. N.* The Nature of Statistical Learning Theory // Springer. — 2000.
2. *Bousquet O.* Statistical Learning Theory. Machine Learning Summer School. — Tuebigen, Germany, 2003. — `http://www.kyb.mpg.de/publication.html?user=bousquet`.

3. *Bousquet O., Boucheron S., Lugosi G.* Introduction to Statistical Learning Theory. Advanced Lectures on Machine Learning // Lecture Notes in Artificial Intelligence. — Vol. 3176. — 2004. — Pp. 169–207.
4. *Tipping M. E.* Sparse Bayesian Learning and the Relevance Vector Machine // Journal of Machine Learning Research. — Vol. 1. — 2001. — Pp. 211–244.
5. *Cucker F., Smale S.* On The Mathematical Foundations of Learning // Bulletin of the American Mathematical Society. — Vol. 39, No 1. — 2001. — Pp. 1–49.
6. *Evgeniou T., Pontil M., Poggio T.* Statistical Learning Theory: A Primer // International Journal of Computer Vision. — Vol. 38, No 1. — 2000. — Pp. 9–13.
7. *Heisele B., Verri A., Poggio T.* Learning and Vision Machines // Proceedings of the IEEE. — Vol. 90, No 7. — 2002. — Pp. 1164–75.

УДК 519.2, 004.93

Фундаментальные аспекты теории статистического оценивания функций и распознавания образов

Э. Фокоуэ

Университет им. Кеттеринга
г. Флинт, Мичиган, США, 48504

Статья представляет собой краткий обзор фундаментальных идей, концепций и результатов теории статистического оценивания функций и распознавания образов. Материал опирается на теорию Вапника–Червоненкиса. Особое внимание уделяется тому, чтобы помочь читателю оценить важность распространения классического закона больших чисел на функциональные пространства и ключевую роль, которую играют такие новые понятия, как принцип минимизации эмпирического риска, состоятельность оценок при построении алгоритмов, обеспечивающих оценку функций с оптимальными свойствами.