4.2 Generalization of Glivenko-Cantelli Theorem

In the previous section, it was observed that the supremum difference between the empirical distribution function of a random variable and its true distribution function has an exponential bound in probability. In this section, the Glivenko-Cantelli Theorem is generalized to probabilities of arbitrary events.

Let $Z_1, \ldots, Z_n$ be iid random vectors in $\mathcal{X} = \mathbb{R}^d$ with probability measure

$$\nu(A) = P(Z_1 \in A) \text{ for all measurable sets } A \subset \mathcal{X}.$$

Analogous to the empirical distribution function, the empirical measure ν_n is defined as

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I(Z_i \in A)$$

Consider $\mathcal{A}$, a collection of subsets of $\mathcal{X}$. The main issue at hand is whether the tail probability of uniform deviations of relative frequencies from probabilities

$$P\left(\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon \right)$$

can be bounded as in the Glivenko-Cantelli theorem for the one-dimensional case. For the proof of Glivenko-Cantelli theorem (Theorem 5), we considered $\{\big(I(z_1 \le z), \ldots, I(z_n \le z)\big) : z \in \mathbb{R}\}$ given $(z_1, \ldots, z_n)$ at the conditioning step and its maximum cardinality as we sweep through $z \in \mathbb{R}$. It is not hard to see that the proof can be extended straightforwardly with consideration of the vector of n membership indicators $\big(I(z_1 \in A), \ldots, I(z_n \in A)\big)$ for sets A in $\mathcal{A}$ given $(z_1, \ldots, z_n) \in \mathcal{X}^n$. For each $(z_1, \ldots, z_n)$, define

$$N_{\mathcal{A}}(z_1, \ldots, z_n) := \Big| \{ \big(I(z_1 \in A), \ldots, I(z_n \in A)\big) : A \in \mathcal{A} \} \Big|.$$

$N_{\mathcal{A}}(z_1, \ldots, z_n)$ stands for the number of distinct vectors of indicators as A runs through all sets in $\mathcal{A}$ given $(z_1, \ldots, z_n)$. In other words, it is the number of different subsets of $(z_1, \ldots, z_n)$ picked out by the sets in $\mathcal{A}$.

Definition 1 (Shatter Coefficient)**.**

$$s(\mathcal{A}, n) := \max_{(z_1, \ldots, z_n) \in \mathcal{X}^n} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

$s(\mathcal{A}, n)$ is called the nth *shatter coefficient* of $\mathcal{A}$.

It means the maximal number of different subsets of n points that can be picked out by the sets in $\mathcal{A}$. Another way to think of $s(\mathcal{A}, n)$ is as a measure of the richness of $\mathcal{A}$. For example, the nth shatter coefficient of $\mathcal{A} = \{(-\infty, z] : \ z \in \mathbb{R}\}$ is

$$s(\mathcal{A}, n) = \max_{(z_1, \ldots, z_n) \in \mathbb{R}^n} \Big| \{ (I(z_1 \le z), \ldots, I(z_n \le z)) : \ z \in \mathbb{R} \} \Big| = n + 1.$$

General properties of $s(\mathcal{A}, n)$ will be mentioned later. The development of this notion of complexity of $\mathcal{A}$ is generally considered as a significant contribution to computational learning theory. With the definition of the nth shatter coefficient of $\mathcal{A}$ and similar arguments as in the proof of the Glivenko-Cantelli Theorem, the following inequality can be established for uniform law of large numbers.

Theorem 6 (Vapnik-Chervonenkis inequality (1971))**.** *For any probability measure ν and class of sets $\mathcal{A}$,*

$$P\left(\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)|>\epsilon\right)\leq 8\,s(\mathcal{A},n)\exp(-n\epsilon^2/32).$$

However, note that the inequality is useful only if $s(\mathcal{A},n)$ grows at a sub-exponential rate.

4.3 V-C Inequality and Classification

Getting back to the initial quest of a probability bound of the departure of the empirical risk of a classifier from its true risk uniformly for $f\in\mathcal{F}$ when $|\mathcal{F}|=\infty$, we relate the V-C inequality to the classification problem. Given $\mathcal{F}=\{f:\mathbb{R}^d\to\{0,1\}\}$, a class of decision rules, take a closer look at the empirical error

$$R_n(f)=\frac{1}{n}\sum_{i=1}^{n}I(f(x_i)\neq y_i).$$

Observe that the misclassification indicator is

$$I(f(x_i)\neq y_i)=I((x_i,y_i)\in A_f),$$

where A_f is the error set of f, i.e. the set of all possible (x,y) misclassified by f,

$$A_f=\{(x,y)\in\mathcal{X}\times\{0,1\}:\ f(x)\neq y\}=\{f^{-1}(1)\times\{0\}\}\cup\{f^{-1}(0)\times\{1\}\}.$$

So,

$$R_n(f)=\frac{1}{n}\sum_{i=1}^{n}I((x_i,y_i)\in A_f)=\nu_n(A_f)$$

with the empirical measure $\nu_n(A)=\frac{1}{n}\sum_{i=1}^{n}I((x_i,y_i)\in A)$ for $A\in\mathcal{X}\times\{0,1\}$. Similarly,

$$R(f)=P(f(X)\neq Y)=P((X,Y)\in A_f)=\nu(A_f).$$

The relationship gives the re-expression of $R_n(f)-R(f)$ as $\nu_n(A_f)-\nu(A_f)$, to which the V-C inequality immediately applies with the induced class of error sets $\mathcal{A}_{\mathcal{F}}:=\{A_f:\ f\in\mathcal{F}\}$. Defining the nth shatter coefficient of $\mathcal{F}$ as $s(\mathcal{F},n):=s(\mathcal{A}_{\mathcal{F}},n)$, we have

Proposition 1.

$$P\left(\sup_{f\in\mathcal{F}}|R_n(f)-R(f)|>\epsilon\right)\leq 8\,s(\mathcal{F},n)\exp\{-n\epsilon^2/32\}.$$

The risk bound in the proposition depends on $s(\mathcal{F},n)$, a combinatorial property of $\mathcal{F}$. In the rest of the section, some combinatorial aspects of $s(\mathcal{F},n)$ will be discussed, and a few classes of decision rules of interest are examined. First consider $\bar{\mathcal{A}}=\{A\times\{0\}\cup A^c\times\{1\}:\ A\in\mathcal{A}\}$ for $\mathcal{A}$, a collection of sets in $\mathbb{R}^d$.

Theorem 7. *For every $n\in\mathbb{N}$, $s(\bar{\mathcal{A}},n)=s(\mathcal{A},n)$.*

Remark 5. In light of the above theorem, to study the properties of $\mathcal{A}_{\mathcal{F}}$ (the class of error sets induced by $\mathcal{F}$), it is sufficient to study those of $\mathcal{A} := \{f^{-1}(1) : f \in \mathcal{F}\}$ for classification.

Proof. We will prove the required equality by showing two inequalities. Recall

$$s(\mathcal{A}, n) = \max_{(z_1,\ldots,z_n)\in\mathcal{X}^n} N_{\mathcal{A}}(z_1, \ldots, z_n).$$

i. $s(\mathcal{A}, n) \leq s(\bar{\mathcal{A}}, n)$:
Fix $(x_1, \ldots, x_n) \in \mathcal{X}^n$. If $A \in \mathcal{A}$ picks out $x_1, \ldots, x_k$, then $\bar{A} = A\times\{0\} \cup A^c\times\{1\}$ picks out $(x_1, 0), \ldots, (x_k, 0)$ from $((x_1, 0), \ldots, (x_n, 0)) \in (\mathcal{X}\times\{0, 1\})^n$. Thus,

$$N_{\mathcal{A}}(x_1, \ldots, x_n) \leq N_{\bar{\mathcal{A}}}((x_1, 0), \ldots, (x_n, 0)).$$

From the inequality,

$$\begin{aligned} s(\mathcal{A}, n) = \max_{(x_1,\ldots,x_n)\in\mathcal{X}^n} N_{\mathcal{A}}(x_1, \ldots, x_n) &\leq \max_{(x_1,\ldots,x_n)\in\mathcal{X}^n} N_{\bar{\mathcal{A}}}((x_1, 0), \ldots, (x_n, 0)) \\ &\leq \max_{(z_1,\ldots,z_n)\in(\mathcal{X}\times\{0,1\})^n} N_{\bar{\mathcal{A}}}(z_1, \ldots, z_n) \\ &= s(\bar{\mathcal{A}}, n). \end{aligned}$$

ii. $s(\bar{\mathcal{A}}, n) \leq s(\mathcal{A}, n)$:
Fix n points $(x_1, 0), \ldots, (x_m, 0), (x_{m+1}, 1), \ldots, (x_n, 1) \in (\mathcal{X}\times\{0, 1\})^n$. If $\bar{A} = A\times\{0\} \cup A^c\times\{1\} \in \bar{\mathcal{A}}$ picks out $(x_1, 0), \ldots, (x_k, 0)$ and $(x_{m+1}, 1), \ldots, (x_{m+l}, 1)$, then $A \in \mathcal{A}$ picks out $(x_1, \ldots, x_k, x_{m+l+1}, \ldots, x_n)$ from $x_1, \ldots, x_n$. Thus,

$$N_{\bar{\mathcal{A}}}((x_1, 0), \ldots, (x_n, 1)) \leq N_{\mathcal{A}}(x_1, \ldots, x_n),$$

which gives $s(\bar{\mathcal{A}}, n) \leq s(\mathcal{A}, n)$.

□