Glivenko-Cantelli : VC dimension and Rademacher Complexity

Youssef Barkaoui

Introduction:

Let $\mathcal{F}$ be a family of integrable functions on some probability space $(\mathcal{X}, \Sigma, P)$.

The goal is to estimate

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right|$$

where $X_1, \ldots, X_n$ i.i.d P.

Empirical CDF

Suppose we want to estimate the CDF

$$g(t) := P[X \leq t].$$

We can use the empirical CDF

$$g_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, t]}(X_i).$$

Want to control

$$\|g_n - g\|_\infty = \sup_t |g_n(t) - g(t)|.$$

This is exactly

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E}f(X) \right|$$

where $\mathcal{F} = \{\mathbf{1}_{(-\infty, t]} : t \in \mathbb{R}\}$.

Glivenko-Cantelli . Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with common cumulative distribution function $F(x)$. Let $\mathbb{F}_n(x)$ be the empirical distribution function based on n observations. Then,

$$P\left(\lim_{n \to \infty} \sup_{x < \infty} |\mathbb{F}_n(x) - F(x)| = 0 \right) = 1,$$

or, equivalently,

$$\lim_{n \to \infty} \|\mathbb{F}_n - F\|_\infty \equiv \lim_{n \to \infty} \sup_x |\mathbb{F}_n(x) - F(x)| = 0 \quad \text{with probability 1.}$$

Lemma. Let F be a (nonrandom) distribution function on $\mathbb{R}$. For each $\epsilon > 0$ there exists a finite partition of the real line of the form $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that for $0 \leq j \leq k-1$

$$F(t_{j+1}) - F(t_j) \leq \epsilon.$$

Proof. Let $\epsilon > 0$ be given. Let $t_0 = -\infty$ and for $j \geq 0$ define

$$t_{j+1} = \sup\{z : F(z) \leq F(t_j) + \epsilon\}.$$

Note that $F(t_{j+1}) \geq F(t_j) + \epsilon$. To see this, suppose that $F(t_{j+1}) < F(t_j) + \epsilon$. Then, by right continuity of F, there would exist $\delta > 0$ so that $F(t_{j+1}+\delta) < F(t_j)+\epsilon$, which would contradict the definition of t_{j+1}. Thus, between t_j and t_{j+1}, F jumps by at least ϵ. Since this can happen at most a finite number of times, the partition is of the desired form, that is $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ with $k < \infty$. Moreover, $F(t_{j+1}^-) \leq F(t_j) + \epsilon$. To see this, note that by definition of t_{j+1} we have $F(t_{j+1} - \delta) \leq F(t_j) + \epsilon$ for all $\delta > 0$. The desired result thus follows from the definition of $F(t_{j+1}^-)$. □

Proof. It suffices to show that for any $\epsilon > 0$

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \epsilon \quad \text{a.s.}$$

To this end, let $\epsilon > 0$ be given and consider a partition of the real line into finitely many pieces of the form $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that for $0 \leq j \leq k-1$

$$F(t_{j+1}^-) - F(t_j) \leq \frac{\epsilon}{2}.$$

The existence of such a partition is ensured by the previous lemma. For any $x \in \mathbb{R}$, there exists j such that $t_j \leq x < t_{j+1}$. For such j,

$$\hat{F}_n(t_j) \leq \hat{F}_n(x) \leq \hat{F}_n(t_{j+1}^-),$$

$$F(t_j) \leq F(x) \leq F(t_{j+1}^-),$$

which implies that

$$\hat{F}_n(t_j) - F(t_{j+1}^-) \leq \hat{F}_n(x) - F(x) \leq \hat{F}_n(t_{j+1}^-) - F(t_j).$$

Furthermore,

$$\hat{F}_n(t_j) - F(t_j) + F(t_j) - F(t_{j+1}^-) \leq \hat{F}_n(x) - F(x),$$

$$\hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + F(t_{j+1}^-) - F(t_j) \geq \hat{F}_n(x) - F(x).$$

By construction of the partition, we have that

$$\hat{F}_n(t_j) - F(t_j) - \frac{\epsilon}{2} \leq \hat{F}_n(x) - F(x),$$

$$\hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + \frac{\epsilon}{2} \geq \hat{F}_n(x) - F(x).$$

The desired result now follows from the Strong Law of Large Numbers.

□

Definition 1. A function class $\mathcal{F}$ is called **Glivenko-Cantelli** if, as $n \to \infty$,

$$\|P_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \to 0 \quad \text{in probability.}$$

In other words, the empirical process converges uniformly to its mean over the class $\mathcal{F}$ in probability.

If $\mathcal{F}$ is too complex or "too big," it may fail to satisfy the Glivenko-Cantelli property (see example below).

In general, highly complex classes of functions or sets will fail to satisfy the Glivenko-Cantelli property. For instance, consider sampling $X_1, \ldots, X_n \sim F$, where F is a continuous distribution over the interval $[0, 1]$. Let $\mathcal{A}$ represent the collection of all subsets of $[0, 1]$ with finitely many elements.

Since the distribution is continuous, we have $\mathbb{F}(A) = 0$ for each $A \in \mathcal{A}$. However, for the finite sample $\{X_1, \ldots, X_n\}$, we observe that $\mathbb{F}_n(A) = 1$ if A contains all the sample points, i.e.,

$$\sup_{A \in \mathcal{A}} |\mathbb{F}_n(A) - \mathbb{F}(A)| = 1,$$

regardless of the sample size n. Consequently, the class of sets $\mathcal{A}$ is not a Glivenko-Cantelli class. Intuitively, the class of sets $\mathcal{A}$ is "too complex".

We want to generalize the theorem for more broad settings and be able to look at sets S other than intervals.

Glivenko-Cantelli. Let $\mathcal{I} = \{(-\infty, a] : a \in \mathbb{R}\}$. Then, as $n \to \infty$,

$$\sup_{S \in \mathcal{I}} \left| \frac{1}{n} \sum_{i=1}^{n} 1_S(X_i) - \mathbb{P}_n(S) \right| \xrightarrow{a.s.} 0.$$

VC dimension

One approach to controll the uniform deviations defined as $\Delta_n(\mathcal{F}) \coloneqq \sup_{f \in \mathcal{F}} |Pf - P_n f|$. of theses distributions is to use the notion of VC dimension.
before going deeper, we should introduce **shattering**.
A set system $(X, \mathcal{S})$ consists of a set X along with a collection of subsets of X. A subset $A \subseteq X$ is *shattered* by $\mathcal{S}$ if each subset of A can be expressed as the intersection of A with a subset in $\mathcal{S}$.

The *VC-dimension* of a set system is the cardinality of the largest subset of X that can be shattered by $\mathcal{S}$.

Sauer-Shelah. If $\mathcal{A}$ has finite VC dimension d, then for $n \geq d$ we have that

$$s(\mathcal{A}, n) \leq \left(\frac{en}{d}\right)^d \leq (n+1)^d.$$

VC Theorem. For any distribution $\mathbb{P}$ and class of sets $\mathcal{A}$, we have that

$$\mathbb{P}\left(\Delta_n(\mathcal{A}) \geq t\right) = \mathbb{P}\left(\sup_x |F_n(x) - F(x)| \geq t\right) \leq 8 s(\mathcal{A}, n) \exp\left(-\frac{nt^2}{32}\right),$$

where $s(\mathcal{A}, n)$ is the shattering coefficient.

The reader can consult [1] for a full proof.

Corollary. If $d < \infty$, then $\Delta_n(\mathcal{A}) \xrightarrow{P} 0, \implies$ Ais Glivenko-Cantelli.

Rademacher Complexity

Definition 2. The **Rademacher Complexity** of a set $\mathcal{A} \subseteq \mathbb{R}^n$ is the quantity

$$R(\mathcal{A}) = \mathbb{E}_\epsilon \left[\sup_{a \in \mathcal{A}} \langle a, \epsilon \rangle \right],$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are i.i.d. Rademacher random variables.

Recall that the **support function** of a set $\mathcal{A}$ is

$$\sigma_{\mathcal{A}}(v) = \sup_{a \in \mathcal{A}} \langle v, a \rangle.$$

So,

$$R(\mathcal{A}) = \mathbb{E}_\epsilon \left[\sigma_{\mathcal{A}}(\epsilon) \right].$$

Definition 3. Consider a sequence X of random independent $(X_1, \ldots, X_n) \subseteq \mathcal{X}$ and a class of functions $\mathcal{F}$ on $\mathcal{X}$. The **Rademacher complexity** of $\mathcal{F}$ is

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right],$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ are i.i.d. Rademacher random variables.

Rademacher Theorem. $\mathrm{E}[\Delta_n(\mathcal{F})] \leq 2\mathcal{R}_n(\mathcal{F})$.

Rademacher complexity can often be bounded above using more geometric characteristics of the function class, such as covering numbers or bracketing numbers. This mirrors the role of VC theory, which connected the problem of uniform convergence to a purely combinatorial property of the set class.

Proof. Let $\{Y_1, \ldots, Y_n\}$ be another independent identically distributed sample. Then,

$$\begin{aligned}
\mathbb{E}[\Delta_n(\mathcal{F})] &= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right| \right] \\
&= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}_Y[f(Y_i)] \right| \right] \\
&= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_Y \left[\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - f(Y_i)) \right] \right| \right] \\
&\leq \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \mathbb{E}_Y \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - f(Y_i)) \right| \right] \\
&\leq \mathbb{E}_{X,Y} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - f(Y_i)) \right| \right].
\end{aligned}$$

Since the distribution of the difference $f(X_i) - f(Y_i)$ is the same as the distribution of $\epsilon_i(f(X_i) - f(Y_i))$ so we obtain,

$$\mathbb{E}[\Delta_n(\mathcal{F})] \leq \mathbb{E}_{X,Y,\epsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i) - f(Y_i))\right|\right]$$

$$\leq \mathbb{E}_{X,Y,\epsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right| + \sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i f(Y_i)\right|\right]$$

$$\leq 2\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right]$$

$$= 2\mathcal{R}_n(\mathcal{F}).$$

□

If the function class is bounded, i.e. for every $f \in \mathcal{F}$ we have that $\|f\|_\infty \leq M$, then by McDiarmid's inequality, $\Delta_n(\mathcal{F})$ is sharply concentrated around its mean, i.e.

$$\mathbb{P}\left(|\Delta_n(\mathcal{F}) - \mathbb{E}[\Delta_n(\mathcal{F})]| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{(2M)^2}\right).$$

By rearranging the inequality from the previous section, we can bound the generalization error $\Delta_n(\mathcal{F})$ with a certain probability.

Let's set the right-hand side of McDiarmid's inequality to a small value δ to get a high-probability bound:

$$2\exp\left(-\frac{nt^2}{(2M)^2}\right) = \delta.$$

Solving for t, we get:

$$t = M\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

Now, we use the fact that $|\Delta_n(\mathcal{F})| \leq |\mathbb{E}[\Delta_n(\mathcal{F})]| + |\Delta_n(\mathcal{F}) - \mathbb{E}[\Delta_n(\mathcal{F})]|$.

We know from the first part that $\mathbb{E}[\Delta_n(\mathcal{F})] \leq 2\mathcal{R}_n(\mathcal{F})$.

We also know from the second part that with probability at least $1-\delta$, the deviation of $\Delta_n(\mathcal{F})$ from its mean is no more than t.

Putting these pieces together gives the final inequality:

$$\Delta_n(\mathcal{F}) \leq 2\mathcal{R}_n(\mathcal{F}) + M\sqrt{\frac{2\ln(2/\delta)}{n}}.$$

In other words, the convergence of **Rademacher Complexity** is a necessary and sufficient condition for F to be a **Glivenko-Cantelli class**.

References

[1] Devroye, L., Györfi, L., Lugosi, G. (2013). A probabilistic theory of pattern recognition (Vol. 31). Springer Science Business Media.

[2] Vershynin, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science (Vol. 47). Cambridge University Press.