Uniform Convergence of the Empirical Distribution and Its Application to the Limiting Case of the Non-Free-Lunch Theorem

Authors: Wang Junyan, Wang Junyan

Date: 2024-01-08T00:00:00+00:00

Abstract

The No Free Lunch (NFL) theorem is an important result in statistical learning theory. Based on Bayesian modeling, one can deduce that the expectation of the loss/utility function is related to the selection of the hypothesis space for the predictive function. If the true predictive function space is considered unknowable, then an arbitrarily chosen hypothesis function space may not necessarily achieve the optimal expectation of the loss function. This paper analyzes the limiting case of the NFL theorem. By utilizing the uniform convergence of distributions—that is, a local form of the Glivenko-Cantelli theorem—it is obtained that in deterministic and non-deterministic prediction problems under certain conditions, when the sample size tends to infinity, the expectation of the loss/utility function is independent of the specific choice of the hypothesis function space. A byproduct of this work is that using the local form of uniform convergence of distributions derived in this paper, one can deduce the uniform convergence of the total variation of distributions. This property was generally considered non-existent previously.

Full Text

Uniform Convergence of Empirical Distribution and Its Application to the Limit Case of the No Free Lunch Theorem

Zhejiang Technical Institute of Economics
Date: January 8, 2024

The No Free Lunch (NFL) theorem is an important result in statistical learning theory (Wolpert, 1992, 1996, 2002). Based on Bayesian modeling, one can derive that the expectation of the loss/utility function depends on the choice of hypothesis space for the predictive function. If the true predictive function

space is considered unknown, then arbitrarily chosen hypothesis function spaces may not necessarily yield optimal expected loss.

This paper analyzes the limit case of the NFL theorem. Using uniform convergence of distributions—a local form of the Glivenko-Cantelli theorem (VI, 1933; Cantelli, 1933; Dvoretzky et al., 1956; Wei, 2008; Mao et al., 2006)—we obtain that in certain deterministic and non-deterministic prediction problems, when the sample size tends to infinity, the expectation of the loss/utility function becomes independent of the specific choice of hypothesis function space. A byproduct of this work is that the local form of uniform convergence of distributions derived herein can be used to obtain uniform convergence in total variation distance, a property previously considered non-existent (Devroye et al., 1990).

1. Three Forms of Uniform Convergence of Empirical Distribution

The Glivenko-Cantelli theorem is one of the fundamental theorems in mathematical statistics theory. It can be stated as follows:

Theorem 1.1. For a cumulative distribution function $F(x)$ defined on a probability space $(X, \mathcal{A}, P)$, and its empirical cumulative distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq x}$, we have:

$$P(\sup_x |F_n(x) - F(x)| > 0) = 0$$

To facilitate subsequent discussion in the No Free Lunch theorem, we also provide a corollary of the Glivenko-Cantelli theorem:

Corollary 1.2. For a cumulative distribution function $F(x)$ defined on a probability space $(X, \mathcal{A}, P)$ with $X \subset \mathbb{R}$, and its empirical cumulative distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq x}$, we have $\forall \Delta > 0, \epsilon > 0$:

$$P\left(\sup_x \left|D_\Delta^+ F_n(x) - D_\Delta^+ F(x)\right| > \epsilon\right) = 0$$

where $D_\Delta^+ f(x) = f(x + \Delta) - f(x)$.

Proof. Let $\epsilon' = \frac{1}{|\Delta|}\epsilon$, and note that:

$$\sup_x \left|D_\Delta^+ F_n(x) - D_\Delta^+ F(x)\right| = \sup_x \left|\frac{F_n(x+\Delta) - F_n(x)}{\Delta} - \frac{F(x+\Delta) - F(x)}{\Delta}\right|$$

$$\leq \frac{\sup_x |F_n(x+\Delta) - F(x+\Delta)|}{|\Delta|} + \frac{\sup_x |F_n(x) - F(x)|}{|\Delta|}$$

If $\sup_x |F_n(x) - F(x)| \leq \epsilon'$ and $\sup_x |F_n(x+\Delta) - F(x+\Delta)| \leq \epsilon'$, then $\sup_x |D_\Delta^+ F_n(x) - D_\Delta^+ F(x)| \leq \epsilon$. By the Glivenko-Cantelli theorem, for any

$\epsilon' > 0$ we have $P(\sup_x |F_n(x) - F(x)| \leq \epsilon') = 1$ and $\lim P(\sup_x |F_n(x) - F(x)| > \epsilon') = 0$. Similarly, $P(\sup_x |F_n(x + \Delta) - F(x + \Delta)| > \epsilon') = 0$. Therefore:

$$\lim P\left(\sup_x |D^+_\Delta F_n(x) - D^+_\Delta F(x)| \leq \epsilon\right) = \lim P\left(\sup_x |D^+_\Delta F_n(x) - D^+_\Delta F(x)| \leq \epsilon, \sup_x |F_n(x) - F(x)| \leq \epsilon', \sup_x |F$$

Thus $\forall \epsilon > 0$, the result holds.

A similar result can be obtained:

Corollary 1.3. For a cumulative distribution function $F(x)$ defined on a probability space $(X, \mathcal{A}, P)$ with $X \subset \mathbb{R}$, and its empirical cumulative distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i \leq x}$, we have $\forall \Delta > 0, \epsilon > 0$:

$$P\left(\sup_x |D_\Delta F_n(x) - D_\Delta F(x)| > \epsilon\right) = 0$$

where $D_\Delta f(x) = f(x + \Delta) - f(x - \Delta)$.

From this theorem, we can derive a conclusion regarding total variation distance:

Proposition 1.4. Let $(X, \mathcal{A}, P)$ be a probability measure space with $X \subset \mathbb{R}$ and bounded, and let $P : \mathcal{A} \to [0, 1]$ be a probability measure with empirical probability measure $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i \in A}$. Then $\forall \upsilon > 0$:

$$P(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \upsilon) = 0$$

Proof. First, cover A with small balls of radius ϵ: $A \subset A_\epsilon = \cup_{i=1}^N B_\epsilon(x_i)$. Since A is totally bounded, $A \subset \bar{A}$ can always be covered by $N < \infty$ small balls. Therefore:

$$|P_n(A) - P(A) - (P_n(A_\epsilon) - P(A_\epsilon))| \leq dx \leq N \cdot \mathrm{vol}(B_\epsilon)$$

Thus:

$$|P_n(A) - P(A)| \leq |P_n(A_\epsilon) - P(A_\epsilon)| + N \cdot \mathrm{vol}(B_\epsilon) \leq \sum_{i=1}^N |P_n(B_\epsilon(x_i)) - P(B_\epsilon(x_i))| + N \cdot \mathrm{vol}(B_\epsilon)$$

$$\leq 2\epsilon \sum_{i=1}^N |D_\epsilon P_n(B_\epsilon(x_i)) - D_\epsilon P(B_\epsilon(x_i))| + N \cdot \mathrm{vol}(B_\epsilon)$$

Therefore:

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \leq \sup_i |D_\epsilon P_n(B_\epsilon(x_i)) - D_\epsilon P(B_\epsilon(x_i))| + N \cdot \mathrm{vol}(B_\epsilon)$$

$$\leq \sup_i |D_\epsilon P_n(B_\epsilon(x_i)) - D_\epsilon P(B_\epsilon(x_i))| + N_{\max} \cdot \mathrm{vol}(B_\epsilon)$$

$$\leq 2\epsilon N_{\max} \sup_i |D_\epsilon P_n(B_\epsilon(x_i)) - D_\epsilon P(B_\epsilon(x_i))| + N_{\max} \cdot \mathrm{vol}(B_\epsilon)$$

where $N_{\max}$ is the maximum number of balls of radius ϵ needed to cover any $A \in \mathcal{A}$. Since $A \subset X$ and X has a finite cover, any A also has a finite cover.

By Corollary 1.3, we know $\forall \epsilon > 0, \varepsilon > 0$: $P(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > N_{\max}(2\epsilon\varepsilon + \mathrm{vol}(B_\epsilon))) = 0$. Let $\upsilon = N_{\max}(2\epsilon\varepsilon + \mathrm{vol}(B_\epsilon))$, which completes the proof.

2. Discussion on the No Free Lunch Theorem

The standard form of the No Free Lunch theorem in learning theory is:

Theorem 2.1. Let C be the learning loss (or utility); let the dataset be $D_m = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$, where x_i are inputs and y_i are outputs; let f represent the input-output relationship function from the true function space, which follows a distribution with probability $P(f)$; and let h represent the input-output relationship function from the solution space, which follows a distribution with probability $P(h)$. Then:

$$E[C|D_m] = \sum_{f,h} E[C|f, h, D_m]p(h|D_m)p(f|D_m)$$

where $E[C|f, h, D_m]$ is the conditional expectation of C.

This theorem represents a Bayesian description of statistical learning problems. It is often interpreted pessimistically: since the distribution $p(f|D_m)$ is unknown, we cannot guarantee that $p(h|D_m)$ determined by the learning algorithm, when weighted with $p(f|D_m)$, will achieve maximal or minimal expected value. However, note that this holds when m is finite. Below we analyze the case when $m \to \infty$.

Deterministic Output

We first discuss a common simple case where $\exists f^* : X \to \mathcal{Y}, y = f^*(x)$. For the learning process, we assume: 1. $h \in H$, where H is the hypothesis space; $f \in F$, where F is the true space; 2. $H_{D_m} = \{h : h(x) = y, \forall (x, y) \in D_m, h \in H\}$, $F_{D_m} = \{f : f(x) = y, \forall (x, y) \in D_m\}$.

Proposition 2.2. $H \supset H_{D_1} \supset H_{D_2} \supset ...$ and $F \supset F_{D_1} \supset F_{D_2} \supset ...$

Furthermore, we can obtain:

Proposition 2.3. Let $(X, \mathcal{A}, P)$ be a probability measure space with cumulative probability distribution $F(x)$ that is Lipschitz continuous, i.e., $\forall x, x' \in X, |F(x) - F(x')| \leq L|x - x'|$ with $L < 1$, and with distinct elements in D_m. Then the following holds with probability 1:

$$P(f(x) \neq f^*(x)|D_m) = 0, \quad P(h(x) \neq f^*(x)|D_m) = 0$$

Proof. Note that $P(f(x) = f^*(x), x \in D_m | D_m) = P(x \in D_m)$. Then:

$$P(f(x) = f^*(x)|D_m) = P(f(x) = f^*(x), x \in D_m|D_m) + P(f(x) = f^*(x), x \notin D_m|D_m) \geq P(x \in D_m)$$

Thus:

$$P(f(x) \neq f^*(x)|D_m) = 1 - P(f(x) = f^*(x)|D_m) \leq 1 - P(x \in D_m) = 1 - \sum_j P(x = x_j)$$

Noting that the x_j are distinct:

$$= 1 - \sum_j [F(x_j + \Delta) - F(x_j)]$$

Consider the following construction: partition X into equal intervals of spacing Δ: $x_1, x_2, ...$ Then:

$$\sum_{i=1}^{|X|/\Delta} [F(x_i + \Delta) - F(x_i)]$$

Note that the above sum excludes the last point in $\{x_i\}$. $\int_X p(x)dx = P(x \in X) = 1$.

By Corollary 1.2, when $m \to \infty$, for any $\epsilon > 0$ and any $[x, x + \epsilon) \subset X$, if $F(x + \epsilon) - F(x) > 0$ then with probability 1 we have $F_m(x + \epsilon) - F_m(x) > 0$. Therefore for any $[x_i, x_i + \epsilon) \subset A$, $P(\exists x' \in D_m : 0 < x' - x_i < \epsilon) = 1$. For $x_1, x_2, ...$, take $x_1', x_2', ...$ such that $0 < x_i' - x_i < \Delta/2$. Then by the Lipschitz condition, with probability 1:

$$\sum_{i=1}^{|X|/\Delta} [F(x_i' + \Delta) - F(x_i')] \geq \sum_{i=1}^{|X|/\Delta} [F(x_i + \Delta) - F(x_i)] - \frac{|X|}{\Delta} \cdot \frac{L\Delta}{2}$$

Since $\{x_i'\} \subset D_m$, we have:

$$\lim_{m \to \infty} \sum_j [F(x_j + \Delta) - F(x_j)] \geq \sum_{i=1}^{|X|/\Delta} [F(x_i' + \Delta) - F(x_i')]$$

Therefore with probability 1:

$$P(f(x) \neq f^*(x)|D_m) \leq 1 - \lim_{m \to \infty} \sum_j [F(x_j + \Delta) - F(x_j)]$$

$$= 1 - \sum_{i=1}^{|X|/\Delta} [F(x_i + \Delta) - F(x_i)] = 1 - P(x \in X) = 0$$

Replacing f with h yields the same result. Hence the proof is complete.

This conclusion shows that when the amount of data tends to infinity, functions in the true space converge in probability to the true solution, and any learning strategy that can guarantee zero learning error also yields an estimated function that converges in probability to the true solution.

From Proposition 2.3, we can further obtain:

Corollary 2.4. Under the same conditions as Proposition 2.3, the following holds with probability 1:

$$P(f(x) \neq h(x)|D_m) = 0$$

Proof. Since $f \neq f^*$ implies $|f - f^*| > 0$, and $h \neq f^*$ implies $|h - f^*| > 0$, and $|f - h| \leq |f - f^*| + |h - f^*|$, by Proposition 2.3:

$$P(|f - h| > 0|D_m) \leq P(|f - f^*| > 0|D_m) + P(|h - f^*| > 0|D_m)$$

Thus:

$$\lim_{m\to\infty} P(|f-h| > 0|D_m) \leq \lim_{m\to\infty} P(|f-f^*| > 0|D_m) + \lim_{m\to\infty} P(|h-f^*| > 0|D_m) = 0$$

This conclusion shows that when the amount of data tends to infinity, any learning strategy that can guarantee zero learning error yields an estimated function that is equal in probability to functions in the true space.

Based on this, we obtain:

Theorem 2.5. Let C be the learning loss (or utility); let the dataset be $D_m = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ where y_i are deterministic outputs; let f represent the input-output relationship function from the true function space following a discrete distribution with probability $P(f)$; let h represent the input-output relationship function from the solution space following a discrete distribution with probability $P(h)$. Assume h and f satisfy the deterministic output hypothesis, and $E(C|f, h, D_m) < \infty$ in NFL Theorem 2.1. Then with probability 1:

$$E_{hf}(C|D_m) = \lim_{m\to\infty} E_{ff}(C|D_m)$$

where $E_{hf}(C|D_m) = E(C|D_m)$ is the left side of NFL Theorem 2.1, and $E_{ff}(C|D_m)$ is the expectation obtained by setting $h = f$ in NFL.

Proof. By Corollary 2.4, the following holds with probability 1:

$$|E_{hf}(C|D_m) - E_{ff}(C|D_m)| = \left| \sum_{h\neq f^*} E_h(h)p(h|D_m) - \sum_{f\neq f^*} E_f(f)p(f|D_m) \right|$$

$$\leq \sum_{h\neq f^*} |E_h(h) - E_h(f^*)|p(h|D_m) + \sum_{f\neq f^*} |E_f(f) - E_f(f^*)|p(f|D_m)$$

Since for discrete $P(f)$ and $P(h)$, $\lim_{m\to\infty} P(h = f^*|D_m) = \lim_{m\to\infty} P(f = f^*|D_m) = 1$, and $\lim_{m\to\infty} P(h \neq f^*|D_m) = \lim_{m\to\infty} P(f \neq f^*|D_m) = 0$, the result follows.

Non-deterministic Output

When no mapping relationship exists between input variable X and output variable Y to be predicted, statistical learning generally employs empirical risk minimization (ERM) to obtain the predictive function. Here we only consider the following scenario:

Definition 2.1. For a probability measure space $((X,Y),\mathcal{A},P)$ and a loss function $l(y,f,Df,...)$ defined on Y and predictive function $f:X\to Y$, $f\in F$ (where $Df...$ denotes derivatives of f), if there exists a unique f^* such that $\forall f\neq f^*$:

$$L_{X,Y}(f^*)=\int l(y,f^*,Df^*,...)dP(X,Y)<\int l(y,f,Df,...)dP(X,Y)=L_{X,Y}(f)$$

then $l(y,f,Df,...)$ and $L_{X,Y}(f)$ are called a regular loss function and regular loss functional, respectively.

Clearly, if $L_{X,Y}(f)$ is strictly convex, then it is a regular loss functional and $l(y,f,Df,...)$ is a regular loss function. However, a regular loss functional is not necessarily strictly convex.

Example 2.1. For $l(y,f)=(y-f(x))^2$, we have:

$$E_{X,Y}[l(f)]=\int(y-f(x))^2dP(X,Y)$$

It is easy to verify that for any $0<\alpha<1$:

$$E_{X,Y}[l(\alpha f_1+(1-\alpha)f_2)]=\int(y-\alpha f_1(x)+(1-\alpha)f_2(x))^2dP(X,Y)$$

$$\leq\int[\alpha(y-f_1(x))^2+(1-\alpha)(y-f_2(x))^2]dP(X,Y)=\alpha E_{X,Y}[l(f_1)]+(1-\alpha)E_{X,Y}[l(f_2)]$$

Thus it is a strictly convex loss functional with a unique optimal solution. Taking the derivative of $E_{X,Y}[l(f)]$ yields:

$$f^*=\int ydP(Y|X)=\arg\min\int(y-f(x))^2dP(X,Y)$$

For regular loss functionals, combined with the uniform convergence property of ERM, we have:

Proposition 2.6. For a probability measure space $((X,Y),\mathcal{A},P)$ and a loss function $l(y,f,Df,...)$ defined on Y and predictive function $f:X\to Y$, $f\in F$, if $\forall\epsilon>0$:

$$P\left(\sup_{f\in F}|L^n_{X,Y}(f)-L_{X,Y}(f)|>\epsilon\right)=0$$

where $L^n_{X,Y}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i), Df(x_i), ...)$ is the empirical estimate of loss functional $L_{X,Y}(f)$. Let $h^* = \arg\min_{f\in F} L^n_{X,Y}(f)$ and $f^* = \arg\min_{f\in F} L_{X,Y}(f)$. Then:

$$P(h^* \neq f^* | D_m) = 0$$

Proof. Suppose $h^* \neq f^*$. Then:

$$L^n_{X,Y}(f^*) - L^n_{X,Y}(h^*) = L^n_{X,Y}(f^*) - L_{X,Y}(f^*) + L_{X,Y}(f^*) - L_{X,Y}(h^*) + L_{X,Y}(h^*) - L^n_{X,Y}(h^*)$$

Note that as $n \to \infty$, $L_{X,Y}(f^*) - L_{X,Y}(h^*) < -3\epsilon$, while:

$$|L^n_{X,Y}(f^*) - L_{X,Y}(f^*)| \leq \sup_{f\in F} |L^n_{X,Y}(f) - L_{X,Y}(f)| \leq \epsilon$$

$$|L_{X,Y}(h^*) - L^n_{X,Y}(h^*)| \leq \sup_{f\in F} |L^n_{X,Y}(f) - L_{X,Y}(f)| \leq \epsilon$$

both hold with probability 1. Therefore $\exists \epsilon > 0$ such that:

$$L^n_{X,Y}(f^*) - L^n_{X,Y}(h^*) \leq -3\epsilon + 2\epsilon = -\epsilon$$

with probability 1. Hence:

$$P(L^n_{X,Y}(h^*) > L^n_{X,Y}(f^*)) = 1$$

But since for any $h^* \neq f^*$, we have $L_{X,Y}(h^*) > L_{X,Y}(f^*)$, and by uniform convergence:

$$P(L^n_{X,Y}(h^*) > L^n_{X,Y}(f^*)) \leq P(L^n_{X,Y}(h^*) \geq L^n_{X,Y}(f^*)) = 0$$

This contradicts the previous result. Therefore, when $n \to \infty$, $h^* = f^*$ always holds. Setting $m = n$ completes the proof.

Similar to Theorem 2.5, we obtain:

Theorem 2.7. Let C be the learning loss (or utility); let the dataset be $D_m = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$ where y_i are non-deterministic outputs; let f represent the input-output relationship function from the true function space following a discrete distribution with probability $P(f)$; let h represent the input-output relationship function from the solution space following a discrete distribution with probability $P(h)$. Assume h and f are obtained by ERM, and $E(C|f, h, D_m) < \infty$ in NFL Theorem 2.1. Then:

$$E_{hf}(C|D_m) = \lim_{m\to\infty} E_{ff}(C|D_m)$$

Proof. This proof is essentially the same as for Theorem 2.5 and is therefore omitted. The difference is that unlike Corollary 2.4, since Proposition 2.6 always holds, this conclusion also always holds.

This paper has analyzed the limit case of the NFL theorem, obtaining the conclusion that when the sample size tends to infinity, the NFL becomes independent of the specific choice of hypothesis space. Previously, due to the NFL theorem, it was believed that ERM-based learning systems could not obtain truly optimal solutions. This paper partially corrects this understanding. The analysis in this work relies on a local form of uniform convergence of distributions (the Glivenko-Cantelli theorem). Based on this, we have obtained uniform convergence in total variation distance, a property previously concluded not to exist. This result has constructive implications for large-sample statistics, data science, and artificial intelligence—fields that rely on massive datasets.

References

茆诗松, 王静龙, 濮晓龙, 2006. 高等数理统计. 第 2 版 [M]. 中国: 高等教育出版社.

韦来生, 2008. 数理统计 (中国科学技术大学数学教学丛书)[M]. 中国: 科学出版社.

Cantelli F P, 1933. Sulla determinazione empirica della leggi di probabilita[J]. Giorn. Ist. Ital. Attuari, 4: 421-424.

Devroye L, Györfi L, 1990. No empirical probability measure can converge in the total variation sense for all distributions[J/OL]. Annals of Statistics, 18: 1496-1499. https://api.semanticscholar.org/CorpusID:15581123.

Dvoretzky A, Kiefer J, Wolfowitz J, 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator[J/OL]. Annals of Mathematical Statistics, 27: 642-669. https://api.semanticscholar.org/CorpusID:122299729.

VI G, 1933. Sulla determinazione empirica della leggi di probabilita[J]. Giorn. Ist. Ital. Attuari, 4: 92-99.

Wolpert D H, 1992. On the connection between in-sample testing and generalization error[J/OL]. Complex Syst., 6. https://api.semanticscholar.org/CorpusID:13901468.

Wolpert D H, 1996. The lack of a priori distinctions between learning algorithms[J/OL]. Neural Computation, 8: 1341-1390. https://api.semanticscholar.org/CorpusID:207609360.

Wolpert D H, 2002. The supervised learning no-free-lunch theorems[M/OL]. London: Springer London: 25-42. https://doi.org/10.1007/978-1-4471-0123-9_3.

Note: Figure translations are in progress. See original paper for figures.