Sorbonne Université
Master Mathématiques et Applications
M2 Mathématiques de la Modélisation

Année 2024/2025
Premier Semestre

On the Nearest Neighbor Method for Regression and Classification

Arnaud Guyader

Contents

Chapter 1

Nearest Neighbor Regression

1.1 Nonparametric regression

Let $(\mathbf{X}, Y)$ be a random couple with values in $\mathbb{R}^d \times \mathbb{R}$. Roughly speaking, the objective of regression analysis is to find a function g such that $Y \approx g(\mathbf{X})$. In this chapter, we will focus our attention on the L_2 risk: we assume that $\mathbb{E}[Y^2] < \infty$ and we want to minimize $\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ when $g : \mathbb{R}^d \to \mathbb{R}$ is Borel[1] and such that $\mathbb{E}\left[g(\mathbf{X})^2\right] < \infty$. This coincides with the notion of conditional expectation.

Definition 1.1 (Regression function)
If $\mathbb{E}[Y^2] < \infty$, $r(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}]$ is the (almost surely) unique random variable such that

$$\mathbb{E}\left[(Y - r(\mathbf{X}))^2\right] = \inf_{g,\ \mathbb{E}[g(\mathbf{X})^2]<\infty} \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right].$$

For any $\mathbf{x} \in \mathbb{R}^d$, the function $x \mapsto \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ is called (a version of) the regression function of Y on $\mathbf{X}$.

In general, since the distribution of $(\mathbf{X}, Y)$ is unknown, the same holds for the regression function. Instead, we suppose that we are given a sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ of i.i.d. random couples with the same distribution as, and independent of, a generic pair $(\mathbf{X}, Y)$. The model may be reformulated as

$$\forall 1 \le i \le n, \qquad Y_i = r(\mathbf{X}_i) + \varepsilon_i,$$

where, by definition of the regression function, the error ε_i satisfies $\mathbb{E}[\varepsilon_i|\mathbf{X}_i] = 0$, and consequently $\mathbb{E}[\varepsilon_i] = 0$. We will adopt the notation

$$\sigma^2(\mathbf{x}) := \mathbb{E}\left[(Y - r(\mathbf{X}))^2|\mathbf{X} = \mathbf{x}\right]$$

for the conditional variance function. Since $\mathbb{E}[(Y - r(\mathbf{X}))^2] < \infty$, we also have $\mathbb{E}[\sigma^2(\mathbf{X})] < \infty$.

For fixed $\mathbf{x} \in \mathbb{R}^d$, our goal is thus to estimate the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ using the data $\mathcal{D}_n$. Such an estimate is denoted $r_n(\mathbf{x}) = r_n(\mathbf{x}, \mathcal{D}_n)$. Hence, for each $\mathbf{x}$, $r_n(\mathbf{x})$ is a random variable, function of $\mathcal{D}_n$, while $r_n(\mathbf{X})$ is a function of $\mathbf{X}$ and $\mathcal{D}_n$.

Definition 1.2 (Consistency)
We say that a regression function estimate $r_n(\mathbf{x})$ is consistent if

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \xrightarrow[n\to\infty]{} 0.$$

It is universally consistent if this holds true for all distributions of $(\mathbf{X}, Y)$ such that $\mathbb{E}[Y^2] < \infty$.

[1] All functions in the present notes are assumed Borel measurable. This will no longer be specified.

Remark. One should keep in mind that in the previous definition the expectation is with respect to $\mathbf{X}$ and $\mathcal{D}_n$.

For the sake of simplicity, we will always work under the following assumption. Nonetheless, all of the subsequent results remain valid without any assumption on the law of $\mathbf{X}$, which we denote by μ. We refer to [2] for the general case as well as for many other results on the nearest neighbor method.

Assumption 1.1 (No mass on hyperspheres)
Assume that $\mathbb{R}^d$ is equipped with the Euclidean norm. For μ-almost every $\mathbf{x} \in \mathbb{R}^d$ and any $r \geq 0$, we suppose that $\mathbb{P}(\|\mathbf{X} - \mathbf{x}\| = r) = \mu(\{\mathbf{x}' \in \mathbb{R}^d,\ \|\mathbf{x}' - \mathbf{x}\| = r\}) = 0$.

In particular, this condition is satisfied as soon as μ is absolutely continuous with respect to Lebesgue's measure on $\mathbb{R}^d$. More importantly, for μ-almost every $\mathbf{x} \in \mathbb{R}^d$, this ensures that we may almost surely reorder the sample as follows[2]

$$\|\mathbf{X}_{(1)}(\mathbf{x}) - \mathbf{x}\| < \cdots < \|\mathbf{X}_{(n)}(\mathbf{x}) - \mathbf{x}\|,$$

and the notation $Y_{(i)}(\mathbf{x})$ stands for the response corresponding to $\mathbf{X}_{(i)}(\mathbf{x})$.

Definition 1.3 (Nearest neighbor estimate)
Let $1 \leq k \leq n$. The k-nearest neighbor estimate of the regression function is defined for any $\mathbf{x} \in \mathbb{R}^d$ by

$$r_n(\mathbf{x}) := \frac{1}{k} \sum_{i=1}^{k} Y_{(i)}(\mathbf{x}).$$

Alternatively, we may also write

$$r_n(\mathbf{x}) := \sum_{i=1}^{n} W_i(\mathbf{x}) Y_i,$$

where $W_i(\mathbf{x}) = 1/k$ if $\mathbf{X}_i$ belongs to the k nearest neighbors of $\mathbf{x}$, and 0 otherwise.

Achtung ! In the sequel, we will be interested in asymptotic properties of $r_n(\mathbf{X})$ when n goes to infinity, and $k = k_n$ will depend on n. However, to lighten the writings, we will usually use the notation k instead of k_n.

1.2 Consistency

Before proceeding with the consistency of the nearest neighbor rule, we start with two technical results. The first one is quite intuitive.

Lemma 1.1 (Convergence of the k-th nearest neighbor)
If $k = k_n = o(n)$, then $\mathbf{X}_{(k)}(\mathbf{X})$ goes in probability to $\mathbf{X}$, i.e. for any $a > 0$,

$$\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{X}) - \mathbf{X}\| \geq a\right) \xrightarrow[n\to\infty]{} 0.$$

Proof. Denote by μ the law of $\mathbf{X}$ on $\mathbb{R}^d$ and recall that the support of μ is defined as

$$\mathcal{S}(\mu) := \left\{\mathbf{x} \in \mathbb{R}^d \text{ such that } \forall \delta > 0,\ \mu(B(\mathbf{x}, \delta)) > 0\right\},$$

[2] Indeed, if $\mathbf{X} \perp\!\!\!\perp \mathbf{X}'$, then by conditioning $\mathbb{P}(\|\mathbf{X} - \mathbf{x}\| = \|\mathbf{X}' - \mathbf{x}\|) = \int \mathbb{P}(\|\mathbf{X} - \mathbf{x}\| = \|\mathbf{x}' - \mathbf{x}\|)\mu(d\mathbf{x}') = 0$.

where $\mu(B(\mathbf{x},\delta))$ stands for the open ball centered at $\mathbf{x}$ and with radius δ. Now, since $\mathbf{X}$ is independent of $\mathbf{X}_1,\ldots,\mathbf{X}_n$, we readily have[3]

$$\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|\geq a\right)=\int_{\mathbb{R}^d}\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|\geq a\right)\mu(dx)=\int_{\mathcal{S}(\mu)}\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|\geq a\right)\mu(dx).$$

Thus, let $\mathbf{x}\in\mathcal{S}(\mu)$ and consider

$$S_n:=\sum_{i=1}^{n}\mathbf{1}_{\|\mathbf{X}_i-\mathbf{x}\|<a}$$

so that

$$\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|\geq a\right)=\mathbb{P}\left(S_n<k\right)=\mathbb{E}\left[\mathbf{1}_{\frac{S_n}{n}-\frac{k}{n}<0}\right].$$

The strong law of large numbers and the fact that $\mathbf{x}$ belongs to the support of μ imply that

$$\frac{S_n}{n}\xrightarrow[n\to\infty]{a.s.}\mu(B(\mathbf{x},a))>0.$$

Since $k/n=k_n/n$ goes to 0 when n goes to infinity, the continuity theorem induces

$$\mathbf{1}_{\frac{S_n}{n}-\frac{k}{n}<0}\xrightarrow[n\to\infty]{a.s.}0.$$

The Lebesgue dominated convergence theorem then gives

$$\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{x})-\mathbf{x}\|\geq a\right)=\mathbb{E}\left[\mathbf{1}_{\frac{S_n}{n}-\frac{k}{n}<0}\right]\xrightarrow[n\to\infty]{}0.$$

Since this is true for any $\mathbf{x}\in\mathcal{S}(\mu)$, another application of the Lebesgue dominated convergence theorem allows us to conclude that

$$\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|\geq a\right)=\int_{\mathcal{S}(\mu)}\mathbb{E}\left[\mathbf{1}_{\frac{S_n}{n}-\frac{k}{n}<0}\right]\mu(dx)\xrightarrow[n\to\infty]{}0.$$

■

Remark. In fact, one can show that $\mathbf{X}_{(k)}(\mathbf{X})$ goes almost surely to $\mathbf{X}$, but convergence in probability is enough for what follows.

The second technical result is due to Stone [7].

Lemma 1.2 (Stone's lemma)
There exists a constant γ_d such that for, any function $\varphi:\mathbb{R}^d\to\mathbb{R}_+$,

$$\mathbb{E}\left[\sum_{i=1}^{n}W_i(\mathbf{X})\varphi(\mathbf{X}_i)\right]\leq\gamma_d\mathbb{E}\left[\varphi(\mathbf{X})\right].$$

Proof. Let $\mathcal{C}=\mathcal{C}(0,\pi/3)$ be a cone with vertex at the origin and angle $\pi/3$. Then, for the standard inner product and the associated Euclidean norm, we have

$$\forall(\mathbf{x},\mathbf{x}')\in\mathcal{C}\times\mathcal{C},\qquad\langle\mathbf{x},\mathbf{x}'\rangle\geq\frac{1}{2}\|\mathbf{x}\|\cdot\|\mathbf{x}'\|.$$

Hence, if $\|\mathbf{x}\|\leq\|\mathbf{x}'\|$, then necessarily $\|\mathbf{x}-\mathbf{x}'\|\leq\|\mathbf{x}'\|$. Next, for any $\mathbf{X}\in\mathbb{R}^d$, it is possible to cover $\mathbb{R}^d$ with a finite number γ_d of cones $\mathcal{C}_j=\mathcal{C}_j(\mathbf{X},\pi/3)$. In each cone $\mathcal{C}_j$, let us mark the k nearest neighbors of $\mathbf{X}$ in that cone[4], in which case we write $\mathbf{X}_i=\mathbf{X}_i^\star$. By the previous argument,

[3] Recall that if Y is independent of $X\sim\mu$, then $\mathbb{E}[\varphi(X,Y)]=\int\mathbb{E}[\varphi(x,Y)]\mu(dx)$.
[4] If there are less than k points $\mathbf{X}_i$'s in a cone, just mark them all.

it turns out that if a point $\mathbf{X}_i$ is not marked, then $\mathbf{X}$ does not belong to the k nearest neighbors of $\mathbf{X}_i$ among $\{\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n\}$. In this respect, let us specify a bit the notation for the weights, namely

$$W_i(\mathbf{X}) = W_i(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n),$$

which is non zero and equal to $1/k$ if and only if $\mathbf{X}_i$ is one of the k nearest neighbors of $\mathbf{X}$ among $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. Thus

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})\varphi(\mathbf{X}_i)\right] = \mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n)\varphi(\mathbf{X}_i)\right] = \sum_{i=1}^{n} \mathbb{E}\left[W_i(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n)\varphi(\mathbf{X}_i)\right].$$

Each expectation might be seen as

$$\mathbb{E}\left[\psi_i(\mathbf{X}; \mathbf{X}_1, \ldots, \mathbf{X}_n)\right] = \mathbb{E}\left[\psi_i(\mathbf{X}_i; \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n)\right],$$

because the $(n+1)$ random variables are i.i.d. (the fact that they are exchangeable suffices to have this equality). Therefore, we may also write

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})\varphi(\mathbf{X}_i)\right] &= \sum_{i=1}^{n} \mathbb{E}\left[W_i(\mathbf{X}_i; \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n)\varphi(\mathbf{X})\right] \\
&= \mathbb{E}\left[\varphi(\mathbf{X}) \sum_{i=1}^{n} W_i(\mathbf{X}_i; \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n)\right].
\end{aligned}$$

The above reasoning simply means that

$$W_i(\mathbf{X}_i; \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n) \leq \frac{1}{k}\mathbf{1}_{\mathbf{X}_i = \mathbf{X}_i^\star},$$

and, consequently,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})\varphi(\mathbf{X}_i)\right] \leq \frac{1}{k}\mathbb{E}\left[\varphi(\mathbf{X}) \sum_{i=1}^{n} \mathbf{1}_{\mathbf{X}_i = \mathbf{X}_i^\star}\right] \leq \gamma_d \mathbb{E}\left[\varphi(\mathbf{X})\right],$$

where the last inequality is due to the fact that there are γ_d cones and, in a given cone, the number of marked points cannot be greater than k.

■

Remark. It can be proved that (see [4], Lemma 5.5)

$$\gamma_d \leq \left(1 + \frac{2}{\sqrt{2 - \sqrt{3}}}\right)^d - 1.$$

We can now proceed with the consistency of the nearest neighbor rule.

Theorem 1.1 (Consistency of the nearest neighbors method)
If $k_n \to \infty$ and $k_n/n \to 0$ when n goes to infinity, then the nearest neighbors method is universally consistent, i.e. for all $(\mathbf{X}, Y)$ such that $\mathbb{E}[Y^2] < \infty$, one has

$$\mathbb{E}\left[\left(r_n(\mathbf{X}) - r(\mathbf{X})\right)^2\right] \xrightarrow[n\to\infty]{} 0.$$

Proof. In order to decompose the L_2 risk into a bias and a variance term, it is natural to introduce

$$\hat{r}_n(\mathbf{x}) := \sum_{i=1}^{n} W_i(\mathbf{x}) r(\mathbf{X}_i) = \frac{1}{k} \sum_{i=1}^{k} r(\mathbf{X}_{(i)}(\mathbf{x})).$$

Just notice that

$$\hat{r}_n(\mathbf{X}) = \mathbb{E}\left[r_n(\mathbf{X}) | \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n\right],$$

so that, by orthogonality,

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] = \mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] + \mathbb{E}\left[(\hat{r}_n(\mathbf{X}) - r(\mathbf{X}))^2\right], \tag{1.1}$$

and the goal is to prove that both terms go to 0. For the variance term, we have

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} W_i(\mathbf{X})(Y_i - r(\mathbf{X}_i))\right)^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} W_i(\mathbf{X})\varepsilon_i\right)^2\right].$$

where $\varepsilon_i = Y_i - r(\mathbf{X}_i)$. Recall that $\mathbb{E}[\varepsilon_i | \mathbf{X}_i] = 0$, so for $i \neq j$,

$$\begin{aligned}
\mathbb{E}\left[W_i(\mathbf{X}) W_j(\mathbf{X}) \varepsilon_i \varepsilon_j\right] &= \mathbb{E}\left[\mathbb{E}\left[W_i(\mathbf{X}) W_j(\mathbf{X}) \varepsilon_i \varepsilon_j | \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i\right]\right] \\
&= \mathbb{E}\left[W_i(\mathbf{X}) W_j(\mathbf{X}) \varepsilon_i \mathbb{E}\left[\varepsilon_j | \mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i\right]\right] \\
&= \mathbb{E}\left[W_i(\mathbf{X}) W_j(\mathbf{X}) \varepsilon_i \mathbb{E}\left[\varepsilon_j | \mathbf{X}_j\right]\right] \\
&= 0.
\end{aligned}$$

This yields

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})^2 \left(Y_i - r(\mathbf{X}_i)\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})^2 \sigma^2(\mathbf{X}_i)\right], \tag{1.2}$$

where the last equality is by conditioning upon $(\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n)$. Then, since $0 \leq W_i \leq 1/k$, we get

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] \leq \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X}) \sigma^2(\mathbf{X}_i)\right]. \tag{1.3}$$

By Lemma 1.2, the variance term satisfies

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] \leq \frac{\gamma_d \mathbb{E}[\sigma^2(\mathbf{X})]}{k},$$

and goes to 0 provided $k = k_n$ goes to infinity. Let us now turn to the bias term, namely

$$\begin{aligned}
\mathbb{E}\left[(\hat{r}_n(\mathbf{X}) - r(\mathbf{X}))^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} W_i(\mathbf{X})(r(\mathbf{X}_i) - r(\mathbf{X}))\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})(r(\mathbf{X}_i) - r(\mathbf{X}))^2\right],
\end{aligned} \tag{1.4}$$

by Jensen's inequality. Now, the set of continuous functions of bounded support is dense in $L_2(\mu)$, see e.g. [5] Theorem A.1. Since $\mathbb{E}[r(\mathbf{X}))^2] < \infty$, this ensures that for any $\varepsilon > 0$, there exists a continuous function r_ε, with compact support, such that

$$\mathbb{E}\left[(r_\varepsilon(\mathbf{X}) - r(\mathbf{X}))^2\right] \leq \varepsilon.$$

Since $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and in view of (1.4), we have

$$\mathbb{E}\left[(\hat{r}_n(\mathbf{X})-r(\mathbf{X}))^2\right] \leq 3\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}_i))^2\right] + 3\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r_\varepsilon(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}))^2\right] + 3\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r_\varepsilon(\mathbf{X})-r(\mathbf{X}))^2\right]. \quad (1.5)$$

The last term is easy, since

$$\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r_\varepsilon(\mathbf{X})-r(\mathbf{X}))^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^n W_i(\mathbf{X})\right)(r_\varepsilon(\mathbf{X})-r(\mathbf{X}))^2\right] = \mathbb{E}\left[(r_\varepsilon(\mathbf{X})-r(\mathbf{X}))^2\right] \leq \varepsilon$$

while for the first one, Stone's Lemma 1.2 gives

$$\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}_i))^2\right] \leq \gamma_d \varepsilon.$$

For the second term, since r_ε is continuous with compact support, it is uniformly continuous. Hence, there exists $a>0$ such that $(r_\varepsilon(\mathbf{x}')-r_\varepsilon(\mathbf{x}))^2 \leq \varepsilon$ as soon as $\|\mathbf{x}'-\mathbf{x}\| \leq a$. Moreover r_ε is bounded, say by C, so

$$(r_\varepsilon(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}))^2 = (r_\varepsilon(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}))^2\mathbf{1}_{\|\mathbf{X}_i-\mathbf{X}\|\leq a} + (r_\varepsilon(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}))^2\mathbf{1}_{\|\mathbf{X}_i-\mathbf{X}\|>a} \leq \varepsilon + 4C^2\mathbf{1}_{\|\mathbf{X}_i-\mathbf{X}\|>a}$$

and

$$\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r_\varepsilon(\mathbf{X}_i)-r_\varepsilon(\mathbf{X}))^2\right] \leq \varepsilon + 4C^2\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})\mathbf{1}_{\|\mathbf{X}_i-\mathbf{X}\|>a}\right].$$

Next, observe that

$$\mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})\mathbf{1}_{\|\mathbf{X}_i-\mathbf{X}\|>a}\right] = \frac{1}{k}\sum_{i=1}^k \mathbb{P}\left(\|\mathbf{X}_{(i)}(\mathbf{X})-\mathbf{X}\|>a\right) \leq \mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|>a\right),$$

and, provided $k=k_n=o(n)$, Lemma 1.1 yields

$$\mathbb{P}\left(\|\mathbf{X}_{(k)}(\mathbf{X})-\mathbf{X}\|>a\right) \xrightarrow[n\to\infty]{} 0.$$

Returning to the bias term (1.5), we have established that

$$\limsup_{n\to\infty} \mathbb{E}\left[(\hat{r}_n(\mathbf{X})-r(\mathbf{X}))^2\right] \leq (2+\gamma_d)\varepsilon.$$

Since $\varepsilon>0$ is arbitrary, the proof is complete.

■

1.3 Rates of convergence

The objective of this section is to go one step further and exhibit rates of convergence for the L_2 error. Under appropriate assumptions, it turns out that, when $d \geq 2$,

$$\mathbb{E}\left[(r_n(\mathbf{X})-r(\mathbf{X}))^2\right] = \mathcal{O}\left(n^{-\frac{2}{d+2}}\right).$$

The upcoming result is the key ingredient to get rates of convergence for the nearest neighbor rule. From now on, we suppose that $\mathbb{R}^d$ is equipped with the **supremum norm** and, in order to avoid any problem with ties, hyperspheres in Assumption 1.1 are replaced with hypercubes. But, again, the upcoming results remain correct without this assumption.

Proposition 1.1 (Rate of convergence of the nearest neighbor)
Let $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_N$ i.i.d. with values in $[0,1]^d$ and denote $\mathbf{X}_{(1,N)}(\mathbf{X})$ the nearest neighbor of $\mathbf{X}$ among $\mathbf{X}_1, \dots, \mathbf{X}_N$, then

- *For $d = 1$,*
$$\mathbb{E}\left[|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}|^2\right] \leq \frac{2}{N+1}.$$
- *For $d \geq 2$,*
$$\mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2\right] \leq 4(N+1)^{-\frac{2}{d}}.$$

Proof. Consider $\mathbf{X}_{N+1}$ with the same distribution as, and independent of $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_N$. Denote $\mathbf{X}_{(1,N)}(\mathbf{X}_i)$ the nearest neighbor of $\mathbf{X}_i$ among $\mathbf{X}_1, \dots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \dots, \mathbf{X}_{N+1}$. By symmetry, we have for all index i
$$\mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2\right] = \mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|^2\right],$$
so
$$\mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2\right] = \frac{1}{N+1}\sum_{i=1}^{N+1}\mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|^2\right]. \tag{1.6}$$
For all $i \in [\![1, N]\!]$, denote $R_i := \|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|$ and B_i the open ball centered at $\mathbf{X}_i$ with radius $R_i/2$. Clearly these balls are all disjoint. By assumption, $R_i \leq 1$ so $\lambda(B_i) \leq 1$. In addition, since $R_i/2 \leq 1/2$ for all i, we also have
$$\bigcup_{i=1}^{N+1} B_i \subset \left[-\frac{1}{2}; \frac{3}{2}\right]^d,$$
so
$$\sum_{i=1}^{N+1}\|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|^d = \sum_{i=1}^{N+1}\lambda(B_i) = \lambda\left(\bigcup_{i=1}^{N+1} B_i\right) \leq \lambda\left([-1/2; 3/2]^d\right) = 2^d. \tag{1.7}$$
If $d \geq 2$, Jensen's inequality for the convex mapping $y \mapsto y^{d/2}$ yields
$$\left(\frac{1}{N+1}\sum_{i=1}^{N+1}\|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|^2\right)^{d/2} \leq \frac{1}{N+1}\sum_{i=1}^{N+1}\|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|^d \leq \frac{2^d}{N+1},$$
so
$$\frac{1}{N+1}\sum_{i=1}^{N+1}\|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i\|^2 \leq \frac{4}{(N+1)^{\frac{2}{d}}},$$
which, coming back to (1.6), yields
$$\mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2\right] \leq \frac{4}{(N+1)^{\frac{2}{d}}}.$$
If $d = 1$, the fact that $0 \leq |\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i| \leq 1$ and (1.7) give
$$\frac{1}{N+1}\sum_{i=1}^{N+1}|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i|^2 \leq \frac{1}{N+1}\sum_{i=1}^{N+1}|\mathbf{X}_{(1,N)}(\mathbf{X}_i) - \mathbf{X}_i| \leq \frac{2}{N+1}.$$
■

Remark. For $d = 1$, one might wonder if the rate

$$\mathbb{E}\left[|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}|^2\right] = \mathcal{O}(N^{-1})$$

is optimal. In fact it is, as can be seen from the following elementary example: consider $m \geq 1$ and $\mathbf{X}$ with density $f(\mathbf{x}) = m\mathbf{x}^{m-1}\mathbf{1}_{[0,1]}(\mathbf{x})$. Then there exists a constant $c = c(m)$ such that, for any $N \geq 1$,

$$\mathbb{E}\left[|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}|^2\right] \geq cN^{-1+\frac{2}{m}}.$$

In particular, this shows that the rate of convergence of the left quantity to zero cannot be faster than $\mathcal{O}(N^{-1})$.

In order to establish rates of convergence, we consider the following framework.

Assumption 1.2

(a) The random variable $\mathbf{X}$ *is bounded, namely* $\mathbf{X} \in [0,1]^d$.

(b) The regression function r *is* L*-Lipschitz with respect to the supremum norm.*

(c) For μ*-almost every* $\mathbf{x} \in \mathcal{S}(\mu)$*, we have* $\sigma^2(\mathbf{x}) \leq \sigma^2 < \infty$.

The main result of this section is

Theorem 1.2 (Rates of convergence)
Under Assumption 1.2, one has:

- *For* $d = 1$,

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \leq \frac{\sigma^2}{k} + 2L^2\left(\frac{k}{n}\right).$$

- *For* $d \geq 2$,

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \leq \frac{\sigma^2}{k} + 4L^2\left(\frac{k}{n}\right)^{\frac{2}{d}}.$$

Proof. We start with the bias-variance decomposition (1.1), namely

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] = \mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] + \mathbb{E}\left[(\hat{r}_n(\mathbf{X}) - r(\mathbf{X}))^2\right].$$

The variance term is straightforward, for by (1.2)

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] = \mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})^2\sigma^2(\mathbf{X}_i)\right].$$

Taking into account that $\sigma^2(\mathbf{x}) \leq \sigma^2$ μ-almost surely, we get

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] \leq \frac{\sigma^2}{k}.$$

The bias term requires more attention. From (1.4), recall that

$$\mathbb{E}\left[(\hat{r}_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \leq \mathbb{E}\left[\sum_{i=1}^n W_i(\mathbf{X})(r(\mathbf{X}_i) - r(\mathbf{X}))^2\right] = \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^k (r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X}))^2\right].$$

Then, the Lipschitz hypothesis yields

$$\mathbb{E}\left[(\hat{r}_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \leq \frac{L^2}{k}\mathbb{E}\left[\sum_{i=1}^k \left\|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\right\|^2\right].$$

In order to upper-bound the last quantity, denote $N := \lfloor n/k \rfloor$ and consider the partitioning

$$\{X_1, \dots, X_n\} = \left(\bigcup_{j=1}^{k} \{X_{(j-1)N+1}, \dots, X_{jN}\} \right) \cup \{X_{kN+1}, \dots, X_n\}.$$

For all $1 \leq j \leq k$, we denote $\mathbf{X}^{(j)}_{(1,N)}(\mathbf{X})$ the nearest neighbor of $\mathbf{X}$ in the j-th subset. Then it is readily seen that

$$\sum_{i=1}^{k} \|\mathbf{X}_{(i)}(\mathbf{X}) - \mathbf{X}\|^2 \leq \sum_{j=1}^{k} \|\mathbf{X}^{(j)}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2.$$

Since the couples $(\mathbf{X}^{(j)}_{(1,N)}(\mathbf{X}), \mathbf{X})_{1 \leq j \leq k}$ have the same law, this leads to

$$\frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{k} \left\| \mathbf{X}_{(i,n)}(\mathbf{X}) - \mathbf{X}) \right\|^2 \right] \leq \mathbb{E}\left[\|\mathbf{X}^{(1)}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2 \right] = \mathbb{E}\left[\|\mathbf{X}_{(1,N)}(\mathbf{X}) - \mathbf{X}\|^2 \right].$$

The conclusion follows from Proposition 1.1, taking into account that $N + 1 = \lfloor n/k \rfloor + 1 \geq n/k$. ■

Thus, balancing bias and variance in the previous result gives the following rates of convergence.

Corollary 1.1
Under Assumption 1.2, one has:

- *For $d = 1$, there exists a sequence (k_n) with $k_n \sim \frac{\sigma\sqrt{n}}{\sqrt{2}L}$ and a universal constant c_1 such that*

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2 \right] \leq c_1 \frac{\sigma L}{\sqrt{n}}.$$

- *For $d \geq 2$, there exists a sequence (k_n) with $k_n \sim \left(\frac{\sigma^2}{4L^2}\right)^{\frac{d}{d+2}} n^{\frac{2}{d+2}}$ and a universal constant c_d such that*

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2 \right] \leq c_d \left(\frac{\sigma^2 L^d}{n} \right)^{\frac{2}{d+2}}.$$

Remark. The rate $n^{-\frac{2}{d+2}}$ illustrates the curse of dimensionality. This phenomenon is made more precise below.

1.4 Further results

1.4.1 Optimality

Let $\mathcal{F}$ be the class of distributions of $(\mathbf{X}, Y)$ that satisfy Assumption 1.2, and $\tilde{r}_n$ any estimator of the regression function. It turns out that the rate of convergence obtained through the nearest neighbor method is optimal in the following sense (see Theorem 3.2 in [5] for a proof of this result that was first established in [8]): for any $d \geq 1$, there exists a constant $\Lambda_d > 0$ such that

$$\liminf_{n \to \infty} \inf_{\tilde{r}_n} \sup_{(X,Y) \in \mathcal{F}} \frac{\mathbb{E}\left[(\tilde{r}_n(\mathbf{X}) - r(\mathbf{X}))^2 \right]}{(\sigma^2 L^d)^{\frac{2}{d+2}} n^{-\frac{2}{d+2}}} \geq \Lambda_d.$$

Hence, for $d \geq 2$, the nearest neighbor regression estimator is minimax (i.e., roughly speaking, the best in the worst case).

1.4.2 Data-splitting

In Corollary 1.1, the parameter $k = k_n$ of the estimate with the optimal rate of convergence depends on the unknown distribution of $(\mathbf{X}, Y)$, especially on the smoothness of the regression function measured by the Lipschitz constant L. In this subsection, we present a data-dependent way of choosing $k = k_n$ and explain why, for bounded Y, the estimate with parameter chosen in such an adaptive way achieves the optimal rate of convergence.

To this aim, we split the sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ in two parts of size $\lfloor n/2 \rfloor$ and $n - \lfloor n/2 \rfloor$, respectively. The first half is denoted by $\mathcal{D}_n^\ell$ (learning set) and is used to construct the k-NN estimate[5] $r_{\lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell) = r_{k,\lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell)$. The second half of the sample, denoted by $\mathcal{D}_n^t$ (testing set), is used to choose k by picking $\hat{k}_n \in K = \{1, \ldots, \lfloor n/2 \rfloor\}$ to minimize the empirical risk

$$\frac{1}{n - \lfloor n/2 \rfloor} \sum_{i=\lfloor n/2 \rfloor + 1}^{n} \left(Y_i - r_{k,\lfloor n/2 \rfloor}(\mathbf{X}_i) \right)^2 .$$

Define the estimate

$$r_n(\mathbf{x}) = r_{\hat{k}_n, \lfloor n/2 \rfloor}(\mathbf{x}, \mathcal{D}_n^\ell),$$

and note that r_n depends on the entire data $\mathcal{D}_n$. If $|Y| \leq M < \infty$ almost surely, a straightforward adaptation of Theorem 7.1 in [5] shows that, for any $\delta > 0$,

$$\mathbb{E}\left[\left(r_n(\mathbf{X}) - r(\mathbf{X}) \right)^2 \right] \leq (1 + \delta) \inf_{k \in K} \mathbb{E}\left[\left(r_{k,\lfloor n/2 \rfloor}(\mathbf{X}) - r(\mathbf{X}) \right)^2 \right] + \lambda \frac{\ln n}{n},$$

for some positive constant λ depending only on M, d and δ. Immediately from Corollary 1.1 we can conclude:

Theorem 1.3
Let $d \geq 2$, suppose that $\mathbf{X}$ has bounded support, $|Y| \leq M$, and r is L-Lipschitz. Let r_n be the k-NN estimate with $k \in K = \{1, \ldots, \lfloor n/2 \rfloor\}$ chosen by data-splitting. If $(\ln n)^{(d+2)/(2d)} n^{-1/2} \leq L$, then

$$\mathbb{E}\left[\left(r_n(\mathbf{X}) - r(\mathbf{X}) \right)^2 \right] \leq \Lambda \left(\frac{L^d}{n} \right)^{\frac{2}{d+2}},$$

for some positive constant Λ which depends only on M, d, and the diameter of the support of $\mathbf{X}$.

Thus, the expected error of the estimate obtained via data-splitting is bounded from above up to a constant by the corresponding minimax lower bound for the class $\mathcal{F}$ of regression functions, with the optimal dependence in L.

1.4.3 Local averaging rules

The nearest neighbor method is an example of local averaging rule. A local averaging estimate of the regression function is an estimate that can be written as

$$r_n(\mathbf{x}) := r_n(\mathbf{x}, \mathcal{D}_n) = \sum_{i=1}^{n} W_{n,i}(\mathbf{x}) Y_i$$

where, for all i, $W_{n,i}(\mathbf{x})$ is a function of $\mathbf{x}$ and $\mathbf{X}_1, \ldots, \mathbf{X}_n$, but not of $Y_1, \ldots, Y_n$. Usually, $W_{n,i}(\mathbf{x})$ is a weight taking values in $[0, 1]$ and such that $\sum_{i=1}^n W_{n,i}(\mathbf{x}) = 1$. As explained before, for the nearest neighbor estimate, the weights are equal to $1/k$ or 0, depending if X_i belongs or not to the k nearest neighbors of $\mathbf{x}$.

[5] For the sake of clarity, we make the dependence of the estimate upon k explicit.

Another example is the so-called kernel estimate. Given a mapping $K : \mathbb{R}^d \to \mathbb{R}_+$, called the kernel, and a bandwith $h > 0$, let

$$W_{n,i}(\mathbf{x}) := \frac{K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right)},$$

so that, for all$\mathbf{x} \in \mathbb{R}^d$, the kernel estimate takes the form

$$r_n(\mathbf{x}) = \frac{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right) Y_i}{\sum_{j=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_j}{h}\right)}.$$

In order to obtain a consistent regression estimate, it is intuitively clear that the weights of the points that are close to $\mathbf{x}$ should be larger, as is the case for nearest neighbor estimation. Accordingly, the bandwith $h = h_n$ will depend on n. Two popular kernels are:

- the naive kernel $K(\mathbf{x}) = \mathbf{1}_{\|\mathbf{x}\| \leq 1}$;
- the Gaussian kernel $K(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$.

Sufficient conditions for the consistency of a local averaging regression estimate are given by a famous result due to Stone [7] (see also [5], Theorem 4.1).

Theorem 1.4 (Stone's Theorem)
Assume that, for any distribution of $\mathbf{X}$ *on* $\mathbb{R}^d$*, the following conditions are satisfied:*

(i) There exists a constant C *such that for any function* $\varphi : \mathbb{R}^d \to \mathbb{R}_+$ *and any* n*,*

$$\mathbb{E}\left[\sum_{i=1}^n |W_{n,i}(\mathbf{X})|\varphi(\mathbf{X}_i)\right] \leq C\ \mathbb{E}[\varphi(\mathbf{X})].$$

(ii) There exists $D \geq 1$ *such that, for all* n*,* $\sum_{i=1}^n |W_{n,i}(\mathbf{X})| \leq D$ *almost surely.*

(iii) For all $a > 0$*,*

$$\sum_{i=1}^n |W_{n,i}(\mathbf{X})|\mathbf{1}_{\|\mathbf{X}_i-\mathbf{X}\|>a} \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

(iv)

$$\sum_{i=1}^n W_{n,i}(\mathbf{X}) \xrightarrow[n\to\infty]{\mathbb{P}} 1.$$

(v)

$$\max_{1\leq i\leq n} W_{n,i}(\mathbf{X}) \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Then the corresponding regression estimate is universally consistent: for all distributions of $(\mathbf{X}, Y)$ *such that* $\mathbb{E}[Y^2] < \infty$*, one has*

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \xrightarrow[n\to\infty]{} 0.$$

Example. For nearest neighbor regression estimate with $k = k_n$, conditions (i) and (iv) are clearly always fulfilled. By Lemma 1.1, (iii) is true provided $k_n = o(n)$. Since $\max_{1\leq i\leq n} W_{n,i}(\mathbf{X}) = 1/k_n$, (v) is satisfied as soon as (k_n) goes to infinity. Last but not least, point (i) is much more involved and corresponds to Stone's Lemma 1.2.

Chapter 2

Nearest Neighbor Classification

2.1 Bayes classifier

In supervised classification (or discrimation, or pattern recognition), we still have $\mathbf{X} \in \mathbb{R}^d$ but this time the response variable Y, also called the label or the class, takes values in $\{0, 1\}$. A classifier is a function $\eta : \mathbb{R}^d \to \{0, 1\}$ and the associated error probability is defined as

$$L(\eta) := \mathbb{P}(\eta(\mathbf{X}) \neq Y).$$

Since $Y \in \{0, 1\}$, the regression function takes the form

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}).$$

It is intuitively clear and not difficult to show that the best classifier is the so-called Bayes classifier, defined for all $\mathbf{x} \in \mathbb{R}^d$ by

$$\eta^\star(\mathbf{x}) := \mathbf{1}_{r(\mathbf{x})>\frac{1}{2}} = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) > \mathbb{P}(Y = 0|\mathbf{X} = \mathbf{x}) \\ 0 & \text{otherwise.} \end{cases}$$

We indeed have the following optimality result.

Lemma 2.1
For any classifier η, one has $L(\eta) \leq L(\eta)$. The quantity

$$L^* = L(\eta^\star) = \inf_{\eta:\mathbb{R}^d\to\{0,1\}} L(\eta)$$

is called the Bayes error.

Proof. We have

$$\mathbb{P}(\eta^\star(\mathbf{X}) = Y|\mathbf{X}) = \mathbb{E}\left[\mathbf{1}_{\eta^\star(\mathbf{X})=Y}|\mathbf{X}\right] = \mathbb{E}\left[\mathbf{1}_{\eta^\star(\mathbf{X})=0}\mathbf{1}_{\eta^\star(\mathbf{X})=Y}|\mathbf{X}\right] + \mathbb{E}\left[\mathbf{1}_{\eta^\star(\mathbf{X})=1}\mathbf{1}_{\eta^\star(\mathbf{X})=Y}|\mathbf{X}\right].$$

Thus,

$$\mathbb{P}(\eta^\star(\mathbf{X}) = Y|\mathbf{X}) = \mathbf{1}_{\eta^\star(\mathbf{X})=0}\mathbb{E}\left[\mathbf{1}_{Y=0}|\mathbf{X}\right] + \mathbf{1}_{\eta^\star(\mathbf{X})=1}\mathbb{E}\left[\mathbf{1}_{Y=1}|\mathbf{X}\right].$$

By definition of η, this reduces to

$$\mathbb{P}(\eta^\star(\mathbf{X}) = Y|\mathbf{X}) = \max(\mathbb{P}(Y = 0|\mathbf{X}), \mathbb{P}(Y = 1|\mathbf{X})).$$

In particular, this implies that, for any classifier η,

$$\mathbb{P}(\eta^\star(\mathbf{X}) = Y|\mathbf{X}) - \mathbb{P}(\eta(\mathbf{X}) = Y|\mathbf{X}) \geq 0,$$

and

$$\mathbb{P}(\eta^\star(\mathbf{X}) = Y) - \mathbb{P}(\eta(\mathbf{X}) = Y) = \mathbb{E}\left[\mathbb{P}(\eta^\star(\mathbf{X}) = Y|\mathbf{X}) - \mathbb{P}(\eta(\mathbf{X}) = Y|\mathbf{X})\right] \geq 0.$$

■

2.2 Consistency

As in the nonparametric regression setting, our goal is to construct a classifier based on a sample $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ of i.i.d. random couples with the same distribution as, and independent of, a generic pair $(\mathbf{X}, Y)$. Given $\mathcal{D}_n$, a classifier $\eta_n(\mathbf{x}) = \eta_n(\mathbf{x}, \mathcal{D}_n)$ has values in $\{0, 1\}$ and assigns to each $\mathbf{x}$ a label 0 or 1. Notice that, in the proof of Lemma 2.1, if we condition by $\mathbf{X}$ and $\mathcal{D}_n$ (which are independent), then we obtain that, almost surely,

$$L^\star = L(\eta^\star) \leq \mathbb{P}(\eta_n(\mathbf{X}) \neq Y | \mathcal{D}_n),$$

hence the following definition.

Definition 2.1
The error probability of a classifier η_n is the random variable

$$L(\eta_n) := \mathbb{P}(\eta_n(\mathbf{X}) \neq Y | \mathcal{D}_n).$$

This classifier is universally consistent if, for any distribution of $(\mathbf{X}, Y)$,

$$\mathbb{E}\left[L(\eta_n)\right] = \mathbb{P}(\eta_n(\mathbf{X}) \neq Y) \xrightarrow[n\to\infty]{} L^*.$$

A standard approach to construct a classifier η_n is to estimate the regression function $r(\mathbf{x}) = \mathbb{P}(Y = 1 | \mathbf{X} = \mathbf{x})$ through some regression estimate r_n and then apply a thresholding rule, that is

$$\eta_n(\mathbf{x}) := \mathbf{1}_{r_n(\mathbf{x}) > \frac{1}{2}}.$$

Example. In particular, this is the case for the nearest neighbor classifier, which consists in a majority vote among the k nearest neighbors of a point $\mathbf{x}$, that is

$$\eta_n(\mathbf{x}) := \begin{cases} 1 & \text{if } \frac{1}{k}\sum_{i=1}^k Y_{(i)}(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

From a general viewpoint, the upcoming result ensures that if r_n is a good regression estimate, then η_n is a good classifier. Recall that μ stands for the law of $\mathbf{X}$ in $\mathbb{R}^d$.

Proposition 2.1
For any classifier η_n based on a regression estimate r_n, one has

$$0 \leq L(\eta_n) - L^* \leq 2\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})||\mathcal{D}_n\right] = 2\int_{\mathbb{R}^d} |r_n(\mathbf{x}) - r(\mathbf{x})|\mu(d\mathbf{x}).$$

As a consequence,

$$0 \leq \mathbb{E}\left[L(\eta_n)\right] - L^* \leq 2\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})|\right].$$

Remark. Since $\|X\|_1 \leq \|X\|_p$ for any $p \geq 1$, one also has

$$0 \leq \mathbb{E}\left[L(\eta_n)\right] - L^* \leq 2\left(\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})|^p\right]\right)^{\frac{1}{p}}. \tag{2.1}$$

Proof. Reasoning as in the proof of Lemma 2.1, we get

$$\mathbb{P}(\eta_n(\mathbf{X}) = Y | \mathbf{X}, \mathcal{D}_n) = \mathbf{1}_{\eta_n(\mathbf{X})=0}\mathbb{P}\left(Y = 0 | \mathbf{X}, \mathcal{D}_n\right) + \mathbf{1}_{\eta_n(\mathbf{X})=1}\mathbb{P}\left(Y = 1 | \mathbf{X}, \mathcal{D}_n\right).$$

Since $(\mathbf{X}, Y)$ is independent of $\mathcal{D}_n$ and taking into account that $\mathbf{1}_{\eta_n(\mathbf{X})=0} = 1 - \mathbf{1}_{\eta_n(\mathbf{X})=1}$, the latter reduces to

$$\mathbb{P}(\eta_n(\mathbf{X}) = Y | \mathbf{X}, \mathcal{D}_n) = \mathbf{1}_{\eta_n(\mathbf{X})=0}(1 - r(\mathbf{X})) + \mathbf{1}_{\eta_n(\mathbf{X})=1} r(\mathbf{X}) = (2r(\mathbf{X}) - 1)\mathbf{1}_{\eta_n(\mathbf{X})=1} + (1 - r(\mathbf{X})).$$

In the same vein, we have

$$\mathbb{P}(\eta^\star(\mathbf{X}) = Y|\mathbf{X}) = \mathbf{1}_{\eta^\star(\mathbf{X})=0}(1 - r(\mathbf{X})) + \mathbf{1}_{\eta^\star(\mathbf{X})=1} r(\mathbf{X}) = (2r(\mathbf{X}) - 1)\mathbf{1}_{\eta^\star(\mathbf{X})=1} + (1 - r(\mathbf{X})),$$

which yields

$$\mathbb{P}(\eta_n(\mathbf{X}) \neq Y|\mathbf{X}, \mathcal{D}_n) - \mathbb{P}(\eta^\star(\mathbf{X}) \neq Y|\mathbf{X}) = (2r(\mathbf{X}) - 1)\left(\mathbf{1}_{\eta^\star(\mathbf{X})=1} - \mathbf{1}_{\eta_n(\mathbf{X})=1}\right),$$

so

$$\mathbb{P}(\eta_n(\mathbf{X}) \neq Y|\mathbf{X}, \mathcal{D}_n) - \mathbb{P}(\eta^\star(\mathbf{X}) \neq Y|\mathbf{X}) = |2r(\mathbf{X}) - 1|\mathbf{1}_{\eta^\star(\mathbf{X})\neq\eta_n(\mathbf{X})} \leq 2\,|r_n(\mathbf{X}) - r(\mathbf{X})|\,, \quad (2.2)$$

because $\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})$ implies $|r_n(\mathbf{X}) - r(\mathbf{X})| \geq |r(\mathbf{X}) - 1/2|$. It remains to integrate both sides with respect to the law of $\mathbf{X}$, that is

$$L(\eta_n) - L^* = \mathbb{P}\left(\eta_n(\mathbf{X}) \neq Y|\mathcal{D}_n\right) - L^* \leq 2\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})||\mathcal{D}_n\right].$$

■

Thus, in order to show that a classifier is consistent, it suffices to show that the associated regression estimate is consistent. This is the case for the nearest neighbor classifier.

Corollary 2.1
If $k_n \to \infty$ and $k_n/n \to 0$ when n goes to infinity, then the nearest neighbors classifier is universally consistent, i.e., for all $(\mathbf{X}, Y)$

$$0 \leq \mathbb{E}\left[L(\eta_n)\right] - L^* \xrightarrow[n\to\infty]{} 0.$$

Proposition 2.1 ensures that

$$0 \leq \mathbb{E}\left[L(\eta_n)\right] - L^* \leq 2\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})|\right] \leq 2\left(\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right]\right)^{\frac{1}{2}}.$$

Since Y is a bounded random variable, it satisfies $\mathbb{E}[Y^2] < \infty$ and one can safely apply Theorem 1.1 to conclude. The forthcoming proof proposes an alternative way of showing this result. More precisely, in the context of supervised classification, the bias term can be handled by an analytic argument, namely Lebesgue's differentiation theorem.

Another Proof. According to Proposition 2.1, we just have to establish that

$$\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})|\right] \xrightarrow[n\to\infty]{} 0.$$

As previously, we introduce

$$\hat{r}_n(\mathbf{x}) := \sum_{i=1}^{n} W_i(\mathbf{x}) r(\mathbf{X}_i) = \frac{1}{k}\sum_{i=1}^{k} r(\mathbf{X}_{(i)}(\mathbf{x})),$$

and apply the triangular inequality to get

$$\mathbb{E}\left[|r_n(\mathbf{X}) - r(\mathbf{X})|\right] \leq \mathbb{E}\left[|r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X})|\right] + \mathbb{E}\left[|\hat{r}_n(\mathbf{X}) - r(\mathbf{X})|\right].$$

For the variance term, Cauchy-Schwarz inequality implies

$$\mathbb{E}\left[|r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X})|\right] \leq \left\{\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right]\right\}^{\frac{1}{2}},$$

and, according to (1.2),

$$\mathbb{E}\left[(r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X}))^2\right] = \mathbb{E}\left[\sum_{i=1}^{n} W_i(\mathbf{X})^2 \left(Y_i - r(\mathbf{X}_i)\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{k} \frac{1}{k^2}\left(Y_{(i)}(\mathbf{X}) - r(\mathbf{X}_{(i)}(\mathbf{X}))\right)^2\right].$$

Since $|Y_{(i)}(\mathbf{X}) - r(\mathbf{X}_{(i)}(\mathbf{X}))| \leq 1$, we conclude that

$$\mathbb{E}\left[|r_n(\mathbf{X}) - \hat{r}_n(\mathbf{X})|\right] \leq \frac{1}{\sqrt{k}}.$$

For the bias term, we have

$$\mathbb{E}\left[|\hat{r}_n(\mathbf{X}) - r(\mathbf{X})|\right] = \mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}(r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X}))\right|\right] \leq \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X})\right|\right].$$

Next, if we set

$$d_{(k+1)}(\mathbf{X}) := \|\mathbf{X}_{(k+1)}(\mathbf{X}) - \mathbf{X}\|,$$

then

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X})\right|\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X})\right|\,\middle|\,\mathbf{X}, d_{(k+1)}(\mathbf{X})\right]\right].$$

Now suppose that, given $\mathbf{X}$ and $d_{(k+1)}(\mathbf{X})$, the random variables $\mathbf{X}'_1, \ldots, \mathbf{X}'_k$ are i.i.d. with common distribution the restriction of μ on the open ball $B(\mathbf{X}, d_{(k+1)}(\mathbf{X}))$, i.e. for any test function φ,

$$\mathbb{E}\left[\varphi(\mathbf{X}')|\mathbf{X}, d_{(k+1)}(\mathbf{X})\right] = \frac{1}{\mu(B(\mathbf{X}, d_{(k+1)}(\mathbf{X})))}\int_{B(\mathbf{X}, d_{(k+1)}(\mathbf{X}))}\varphi(\mathbf{x}')\mu(d\mathbf{x}').$$

It is intuitively clear (but a bit tedious to justify, see Lemma A.1 in [3]) that for any test function $\varphi : \mathbb{R}^k \to \mathbb{R}$ that is symmetric in its variables, one has

$$\mathbb{E}\left[\varphi(\mathbf{X}_{(1)}(\mathbf{X}), \ldots, \mathbf{X}_{(k)}(\mathbf{X}))|\mathbf{X}, d_{(k+1)}(\mathbf{X})\right] = \mathbb{E}\left[\varphi(\mathbf{X}'_1, \ldots, \mathbf{X}'_k)|\mathbf{X}, d_{(k+1)}(\mathbf{X})\right].$$

In our case, this gives

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X})\right|\,\middle|\,\mathbf{X}, d_{(k+1)}(\mathbf{X})\right] = \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}'_i) - r(\mathbf{X})\right|\,\middle|\,\mathbf{X}, d_{(k+1)}(\mathbf{X})\right].$$

Given $\mathbf{X}$ and $d_{(k+1)}(\mathbf{X})$, the random variables $\mathbf{X}'_1, \ldots, \mathbf{X}'_k$ are distributed as a generic variable $\mathbf{X}'$ whose law depends on $\mathbf{X}$ and $d_{(k+1)}(\mathbf{X})$, so

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}'_i) - r(\mathbf{X})\right|\,\middle|\,\mathbf{X}, d_{(k+1)}(\mathbf{X})\right] = \mathbb{E}\left[\left|r(\mathbf{X}') - r(\mathbf{X})\right|\middle|\,\mathbf{X}, d_{(k+1)}(\mathbf{X})\right],$$

which finally yields

$$\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\left|r(\mathbf{X}_{(i)}(\mathbf{X})) - r(\mathbf{X})\right|\right] = \mathbb{E}\left[\left|r(\mathbf{X}') - r(\mathbf{X})\right|\right],$$

with

$$\mathbb{E}\left[\left|r(\mathbf{X}') - r(\mathbf{X})\right|\right] = \mathbb{E}\left[\frac{1}{\mu(B(\mathbf{X}, d_{(k+1)}(\mathbf{X})))}\int_{B(\mathbf{X}, d_{(k+1)}(\mathbf{X}))}\left|r(\mathbf{x}') - r(\mathbf{X})\right|\mu(d\mathbf{x}')\right] =: \mathbb{E}\left[I_n\right].$$

By Lebesgue's differentiation theorem, if μ is a σ-finite measure on $\mathbb{R}^d$ that is bounded on compact sets, and if φ is locally integrable with respect to μ, then for μ-almost all $\mathbf{x}$[1],

$$\frac{1}{\mu(B(\mathbf{x}, \delta))}\int_{B(\mathbf{x}, \delta)}\left|\varphi(\mathbf{x}') - \varphi(\mathbf{x})\right|\mu(d\mathbf{x}') \xrightarrow[\delta \to 0]{} 0,$$

[1]Such points $\mathbf{x}$ are sometimes called Lebesgue points.

or, equivalently since $\mathbf{X} \sim \mu$,

$$\frac{1}{\mu(B(\mathbf{X},\delta))} \int_{B(\mathbf{X},\delta)} \left|\varphi(\mathbf{x}') - \varphi(\mathbf{X})\right| \mu(d\mathbf{x}') \xrightarrow[\delta\to 0]{a.s.} 0,$$

which implies that

$$\sup_{0\leq\delta\leq\delta_0} \frac{1}{\mu(B(\mathbf{X},\delta))} \int_{B(\mathbf{X},\delta)} \left|\varphi(\mathbf{x}') - \varphi(\mathbf{X})\right| \mu(d\mathbf{x}') \xrightarrow[\delta_0\to 0]{a.s.} 0. \tag{2.3}$$

In our case, for any $\delta_0 > 0$,

$$\mathbb{E}\left[\left|r(\mathbf{X}') - r(\mathbf{X})\right|\right] = \mathbb{E}\left[I_n \mathbf{1}_{d_{(k+1)}(\mathbf{X})>\delta_0}\right] + \mathbb{E}\left[I_n \mathbf{1}_{d_{(k+1)}(\mathbf{X})\leq\delta_0}\right],$$

and since $0 \leq I_n \leq 1$ for all n, we are led to

$$\mathbb{E}\left[\left|r(\mathbf{X}') - r(\mathbf{X})\right|\right] \leq \mathbb{P}\left(d_{(k+1)}(\mathbf{X}) > \delta_0\right) + \mathbb{E}\left[\sup_{0\leq\delta\leq\delta_0} \frac{1}{\mu(B(\mathbf{X},\delta))} \int_{B(\mathbf{X},\delta)} \left|r(\mathbf{x}') - r(\mathbf{X})\right| \mu(d\mathbf{x}')\right].$$

Thanks to Lemma 1.1, we know that, provided $k = k_n = o(n)$,

$$\mathbb{P}\left(d_{(k+1)}(\mathbf{X}) > \delta_0\right) \xrightarrow[n\to\infty]{} 0,$$

which implies that, for any $\delta_0 > 0$,

$$\limsup_{n\to\infty} \mathbb{E}\left[\left|r(\mathbf{X}') - r(\mathbf{X})\right|\right] \leq \mathbb{E}\left[\sup_{0\leq\delta\leq\delta_0} \frac{1}{\mu(B(\mathbf{X},\delta))} \int_{B(\mathbf{X},\delta)} \left|r(\mathbf{x}') - r(\mathbf{X})\right| \mu(d\mathbf{x}')\right].$$

Then, (2.3) and Lebesgue's dominated convergence theorem ensure that

$$\mathbb{E}\left[\sup_{0\leq\delta\leq\delta_0} \frac{1}{\mu(B(\mathbf{X},\delta))} \int_{B(\mathbf{X},\delta)} \left|r(\mathbf{x}') - r(\mathbf{X})\right| \mu(d\mathbf{x}')\right] \xrightarrow[\delta_0\to 0]{} 0,$$

which completes the proof.

■

2.3 Further results

Assume that the classifier is constructed by thresholding a regression estimator r_n, meaning that

$$\eta_n(\mathbf{x}) = \mathbf{1}_{r_n(\mathbf{x})>\frac{1}{2}}.$$

In order to establish rates of convergence for the classification error, we could thus simply start from rates of convergence for the regression estimator and apply the upper-bound of equation (2.1). Nonetheless, the upcoming result shows that this is not accurate.

Proposition 2.2
Suppose that r_n is universally consistent, i.e.,

$$\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right] \xrightarrow[n\to\infty]{} 0,$$

and that $\eta_n(\mathbf{x}) = \mathbf{1}_{r_n(\mathbf{x})>\frac{1}{2}}$, then

$$\frac{\mathbb{E}\left[L(\eta_n)\right] - L^*}{\sqrt{\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right]}} \xrightarrow[n\to\infty]{} 0.$$

Proof. By integrating (2.2), we get

$$\Delta := \mathbb{E}\left[L(\eta_n)\right] - L^* = 2\mathbb{E}\left[\left|r(\mathbf{X}) - \frac{1}{2}\right| \mathbf{1}_{\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})}\right].$$

Let $\varepsilon > 0$. Since $\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})$ implies $\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq |r(\mathbf{X}) - r_n(\mathbf{X})| \, \mathbf{1}_{r(\mathbf{X}) \neq \frac{1}{2}}$, we may write

$$\mathbb{E}\left[\left|r(\mathbf{X}) - \frac{1}{2}\right| \mathbf{1}_{\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})}\right] \leq \mathbb{E}\left[|r(\mathbf{X}) - r_n(\mathbf{X})| \, \mathbf{1}_{\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})} \mathbf{1}_{r(\mathbf{X}) \neq \frac{1}{2}}\right].$$

Now, the right-hand side is equal to

$$\mathbb{E}\left[|r(\mathbf{X}) - r_n(\mathbf{X})| \, \mathbf{1}_{\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})} \mathbf{1}_{\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq \varepsilon} \mathbf{1}_{r(\mathbf{X}) \neq \frac{1}{2}}\right] + \mathbb{E}\left[|r(\mathbf{X}) - r_n(\mathbf{X})| \, \mathbf{1}_{\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})} \mathbf{1}_{\left|r(\mathbf{X}) - \frac{1}{2}\right| > \varepsilon}\right]$$

and Cauchy-Schwarz inequality gives

$$\Delta \leq \|r_n - r\|_2 \left\{ \mathbb{P}\left(\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq \varepsilon, r(\mathbf{X}) \neq \frac{1}{2}\right)^{\frac{1}{2}} + \mathbb{P}\left(\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X}), \left|r(\mathbf{X}) - \frac{1}{2}\right| > \varepsilon\right)^{\frac{1}{2}} \right\}.$$

For the first term, $\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X})$ and $|r(\mathbf{X}) - \frac{1}{2}| > \varepsilon$ imply $|r_n(\mathbf{X}) - r(\mathbf{X})| > \varepsilon$ so

$$\mathbb{P}\left(\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X}), \left|r(\mathbf{X}) - \frac{1}{2}\right| > \varepsilon\right) \leq \mathbb{P}\left(|r_n(\mathbf{X}) - r(\mathbf{X})| > \varepsilon\right) \leq \frac{\|r_n - r\|_2^2}{\varepsilon^2},$$

and the consistency of r_n imposes

$$\mathbb{P}\left(\eta^\star(\mathbf{X}) \neq \eta_n(\mathbf{X}), \left|r(\mathbf{X}) - \frac{1}{2}\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

Hence, for any $\varepsilon > 0$, we are led to

$$\limsup_{n\to\infty} \frac{\mathbb{E}\left[L(\eta_n)\right] - L^*}{\sqrt{\mathbb{E}\left[(r_n(\mathbf{X}) - r(\mathbf{X}))^2\right]}} = \limsup_{n\to\infty} \frac{\Delta}{\|r_n - r\|_2} \leq \mathbb{P}\left(\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq \varepsilon, r(\mathbf{X}) \neq \frac{1}{2}\right)^{\frac{1}{2}}.$$

It remains to apply Lebesgue dominated convergence theorem to conclude:

$$\mathbb{P}\left(\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq \varepsilon, r(\mathbf{X}) \neq \frac{1}{2}\right) = \mathbb{E}\left[\mathbf{1}_{\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq \varepsilon, r(\mathbf{X}) \neq \frac{1}{2}}\right] \xrightarrow[\varepsilon\to 0]{} 0.$$

■

The previous result is not surprising and simply means that classification is easier than regression (see also [4], Section 6.7). Indeed, in order to obtain a good classifier, it suffices for $\eta_n(\mathbf{x})$ to be on the same side of $1/2$ as $\eta^\star(\mathbf{x})$, whereas a good regression estimate r_n has to be close to r everywhere with respect to the law of $\mathbf{X}$.

In this respect, if $\eta_n = \mathbf{1}_{r_n > \frac{1}{2}}$ with r_n an estimator of r, then the points that are difficult to classify are those for which $r(\mathbf{x}) \approx \frac{1}{2}$. This point can be quantified through the so-called margin condition, which assumes that there exist $c > 0$, $\alpha > 0$, and $0 < t_0 \leq 1/2$ such that, for all $t \in [0, t_0]$,

$$\mathbb{P}\left(\left|r(\mathbf{X}) - \frac{1}{2}\right| \leq t\right) \leq c \, t^\alpha.$$

In particular, one may notice that this implies $\mathbb{P}(r(\mathbf{X}) = \frac{1}{2}) = 0$. More on this topic can be found for example in [9, 1, 6].

Bibliography

[1] Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[2] Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

[3] Frédéric Cérou and Arnaud Guyader. Nearest neighbor classification in infinite dimension. *ESAIM Probability and Statistics*, 10: 340–355, 2006.

[4] László Györfi, Luc Devroye, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[5] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[6] Michael Kohler and Adam Krzyżak. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Trans. Inform. Theory*, 53(5):1735–1742, 2007.

[7] Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–645, 1977.

[8] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

[9] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.