Institute of Mathematics, EPFL MATH524

Chapter 7: Data-dependent partitioning estimators

Lecturer: Rajita Chandak *Spring 2025*

We have already seen a type of partitioning estimator in the form of splines. Splines use a pre-determined scheme to partition the support space of the data and then generate estimates of the regression function based on the data in each resulting partition. We can think of many other modern estimators that have similar underlying structure of partitioning the space in order to generate estimates that can adapt to different features of the regression function that depend on the input. For example, decision trees and neural networks are very popular modern estimation tools that are based on partitioning schemes. And so, it may be beneficial to identify a general result that can be used for establishing consistency of partitioning-based estimators.

In this chapter we will focus only on consistency results. As it turns out minimax or uniform results may not hold for many such estimators. In fact, in some cases (like with neural networks), this remains an open question and an active area of research.

1 A general consistency result

Lets consider the class of functions $\mathcal{G}$ with functions $g : \mathbb{R}^d \to \mathbb{R}$ and partition $\mathcal{P}$ such that

$$\mathcal{G} \circ \mathcal{P} = \left\{ f : \mathbb{R}^d \to R | f = \sum_{A \in \mathcal{P}} g_A \mathbf{1}(A) \text{ for some } g_A \in \mathcal{G}(A \in \mathcal{P}) \right\}.$$

Here, each function of $\mathcal{G} \circ \mathcal{P}$ is obtained by applying a different function of $\mathcal{G}$ in each set in $\mathcal{P}$. Lets start with the simple case of $\mathcal{G} = \mathcal{G}_c$, where $\mathcal{G}_c$ is the set of all constant functions. The least-squares estimator, given by the function f such that

$$\min_{f \in \mathcal{G}_c \circ \mathcal{P}_n} \frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2. \tag{1}$$

One can check that the least-squares estimator is given by

$$\hat{m}_n(\mathbf{x}) = \frac{\sum_i Y_i \mathbf{1}(\mathbf{X}_i \in A(\mathbf{x}))}{\sum_i \mathbf{1}(\mathbf{X}_i \in A(\mathbf{x}))}, \tag{2}$$

where $A(\mathbf{x})$ is the cell of the partition $A \in \mathcal{P}_n$ for which $\mathbf{x} \in A$. Furthermore, note that $\hat{m}_n \in \mathcal{G}_c \circ \mathcal{P}$ since it is the sample average of the data within each cell of the partition, which is constant given the data.

We will be interested specifically in estimators that use a data-dependent partitioning scheme. The reason this specific class of estimators is interesting to consider is that the construction of the

estimator requires using the data (at least) twice. First, the data is used to generate some partition $\mathcal{P}_n = \mathcal{P}_n(\mathcal{D}_n) \in \mathbb{R}^d$, where $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Then, the partition is used to define the estimator $\hat{m}_n(\mathbf{x})$ by applying some pre-determined mapping to the Y_i for which the corresponding $\mathbf{X}_i$ fall into the same partition as $\mathbf{x}$. If the function is a simple averaging of the data in the partition, the estimator is precisely the $\hat{m}_n$ estimator defined earlier. More complicated operations on the data will lead to different function classes $\mathcal{G} \circ \mathcal{P}$ and thus lead to different least-squares estimators. Importantly, by interpreting these estimators as least-squares estimators for different function classes, we can hope to use the tools and theory we have established previously for consistency. But in order to use these methods, we need to work with a truncation of the estimator:

Definition 1 (Truncation)**.** *Let $\beta_n \in \mathbb{R}_+$ such that $\beta_n \to \infty$ as $n \to \infty$. Then define the truncation*

$$m_n(x) = T_{\beta_n}(\hat{m}_n(x)) \tag{3}$$

where

$$T_L(u) = \begin{cases} u & \text{if } |u| \leq L \\ L\text{sign}(u) & \text{otherwise.} \end{cases}$$

Why truncate? Recall the proof of the ULLN in Chapter 5. Truncation allowed for the use of concentration inequalities that involve covering numbers and VC dimension, which we can hope to analyze for a data-dependent partitioning scheme and as a result hopefully also prove consistency of such estimators.

Next, we introduce the partitioning number. This will help relate the VC dimension to the random partitioning scheme involved in constructing $\hat{m}_n$

Definition 2 (Partitioning number)**.** *Let Π be a family of partitions of $\mathbb{R}^d$. For a set of $\mathbf{x}_1^n = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$, let $\Delta(\mathbf{x}_1^n, \Pi)$ be the number of distinct partitions of $\mathbf{x}_1^n$ induced by elements of Π, i.e., $\Delta(\mathbf{x}_1^n, \Pi)$ is the number of different partitions $\{\mathbf{x}_1^n \cap A : A \in \mathcal{P}\}$ of $\mathbf{x}_1^n$ for $\mathcal{P} \in \Pi$. The partitioning number $\Delta_n(\Pi)$ is defined by*

$$\Delta_n(\Pi) = \max\left\{\Delta(\mathbf{x}_1^n, \Pi) : x_1, \ldots, x_n \in \mathbb{R}^d\right\}.$$

In other words, the partitioning number is the maximum number of different partitions of any n point set that can be induced by members of Π.

Recall the examples from Chapter 5 when we defined VC dimension. The difference here is that we consider Π to be a particular collection of partitions (that may be dependent on other features of the problem) instead of considering all possible sets in a class. The VC dimension of a set allows the elements of the set to have non-zero intersections while the partitioning number only considers sets that form a valid partition of the space.

Example 1. *Let Π_k be the collection of all partitions of $\mathbb{R}$ into k intervals. Then, a partition of $\mathbf{x}_1^n$ induced by an element of Π_k is given by some numbers $0 \leq i_1 \leq \ldots \leq i_{k-1} \leq n$ such that the partition is given by*

$$\{x_1, \ldots x_{i_1}\}, \{x_{i_1+1}, \ldots, x_{i_2}\}, \ldots, \{x_{i_{k-1}+1}, \ldots, x_n\}.$$

A counting argument is then needed to identify the partitioning number here. There are a total of $\binom{(n+1+(k-1)-1)}{k-1} = \binom{n+k-1}{n}$ *such tuples for splitting* $\mathbf{x}_1^n$*. Thus,*

$$\Delta(\mathbf{x}_1^n, \Pi_k) = \binom{n+k-1}{n}.$$

Let Π be a family of finite partitions of $\mathbb{R}^d$. Then the maximum number of sets contained in a partition $\mathcal{P} \in \Pi$ is defined as

$$M(\Pi) = \max\{|\mathcal{P}| : \mathcal{P} \in \Pi\}.$$

Furthermore, set

$$\Pi_n = \{\mathcal{P}_n(\{x_i, y_i\}_{i=1}^n) : \{x_i, y_i\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}\}$$

to be the collection of partitions that contain all the data-dependent partitions $\mathcal{P}_n$. We now state one of the main result for establishing consistency of partitioning estimators under fairly general conditions.

Theorem 1

Let m_n *be defined by* (2) *and* (3)*. Assume* $\beta_n \to \infty$ *as* $n \to \infty$*,*

$$\frac{M(\Pi_n)\beta_n^4 \log(\beta_n)}{n} \to 0, \tag{4}$$

$$\frac{\log(\Delta_n(\Pi_n))\beta_n^4}{n} \to 0, \tag{5}$$

$$\frac{\beta_n^4}{n^{1-\eta}} \to 0$$

for some $\eta > 0$*, all as* $n \to \infty$*. Furthermore, suppose*

$$\inf_{S : S \subseteq \mathbb{R}^d, \mu(S) \geq 1-\delta} \mu(\{x : diam\,(A_n(x) \cap S) > \gamma\}) \to 0 \; a.s. \tag{6}$$

for all $\gamma > 0$*,* $\delta \in (0, 1)$*. Then,*

$$\int |m_n(x) - m(x)|^2 \mu(\mathrm{d}x) \to 0 \; a.s.$$

Note that assumption (4) implies that the maximum number of cells in the partition does not grow too fast. Assumption (5) requires the the log-partitioning number is small compared to the sample size and Assumption (6) ensures that the size of the cells converges to zero in some sense.

We will now work towards proving this result. First, we identify that the proof of the theorem implicitly will require the proof of two related statements:

$$\inf_{f \in T_{\beta_n}(\mathcal{G}_c \circ \mathcal{P}_n)} \int (f(x) - m(x))^2 \mu(\mathrm{d}x) \to 0 \tag{7}$$

and

$$\sup_{f \in T_{\beta_n}(\mathcal{G}_c \circ \mathcal{P}_n)} \left| \frac{1}{n} \sum_i |f(X_i) - Y_{i,L}|^2 - \mathbb{E}[|f(X) - Y_L|^2] \right| \to 0 \tag{8}$$

for all $L > 0$ where $Y_L = Y\mathbf{1}(|Y| \leq L)$ and $Y_{i,L} = Y_i\mathbf{1}(|Y_i| \leq L)$.

Note that in (8), we have used the observation that since $\mathcal{G}_c \circ \mathcal{P}_n$ is a collection of piecewise constant functions, $T_{\beta_n}(\mathcal{G}_c \circ \mathcal{P}_n)$ is the collection of functions from $\mathcal{G}_c \circ \mathcal{P}_n$ that are bounded in absolute value by β_n.

The reduction of Theorem 1 to proving (7) and (8) is due to the following result that we will take as given for this course. Motivation for why we would expect this result to hold will be worked through in the exercises.

Theorem 2

Let $\mathcal{F}(\mathcal{D}_n)$ be a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ and consider the estimator that satisfies (1) and (3). If $\beta_n \to \infty$,

$$\lim_{n\to\infty} \inf_{f \in \mathcal{F}_n, \|f\|_\infty \leq \beta_n} \int |f(\mathbf{x}) - m(\mathbf{x})|^2 \mu(\mathrm{d}\mathbf{x}) = 0,$$

and

$$\lim_{n\to\infty} \sup_{f \in T_{\beta_n}(\mathcal{F}_n)} \left| \frac{1}{n} \sum_{j=1}^{n} |f(\mathbf{X}_j) - Y_{j,L}|^2 - \mathbb{E}[|f(\mathbf{X}) - Y_L|^2] \right| = 0$$

for all $L > 0$, then

$$\lim_{n\to\infty} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(\mathrm{d}x) = 0.$$

The same holds for equivalent statements with expectations.

Before proving Theorem 1, we will note the following result that establishes a relationship between covering numbers and partitioning numbers.

Lemma 1

Let $1 \leq p < \infty$ and Π be a family of partitions of $\mathbb{R}^d$. Let $\mathcal{G}$ be a class of functions for $g : \mathbb{R}^d \to \mathbb{R}$. Then, for each $\mathbf{x}_1^n \in \mathbb{R}^d$ and $\varepsilon > 0$,

$$N(\varepsilon, \mathcal{G} \circ \Pi, L_p(\mathbf{x}_1^n)) \leq \Delta(\mathbf{x}_1^n, \Pi) \left(\sup_{\mathbf{z}_1^m \in \mathbf{x}_1^n, m \leq n} N(\varepsilon, \mathcal{G}, L_p(\mathbf{z}_1^m)) \right)^{M(\Pi)}$$

Proof of Lemma 1. For simplicity, lets denote

$$N = \sup_{\mathbf{z}_1^m \in \mathbf{x}_1^n, m \leq n} N(\varepsilon, \mathcal{G}, L_p(\mathbf{z}_1^m)).$$

Now, fix $x_1, \dots, x_n$ and $\varepsilon > 0$. Let $\mathcal{P} = \{A_j\} \in \Pi$ be an arbitrary partition. Then, we can identify the partition of $\mathbf{x}_1^n$ as the collection of sets $B_j = \mathbf{x}_1^n \cap A_j$. For each j, we can choose an ε-cover that is no larger than N (this is always possible by definition of N). This means, for each j, we have a set $\mathcal{G}_{B_j}$ of functions such that for each $g \in \mathcal{G}$, there exists a $\bar{g} \in \mathcal{G}_{B_j}$ that satisfies

$$\frac{1}{|B_j|} \sum_{x \in B_j} |g(x) - \bar{g}(x)|^p \leq \varepsilon^p.$$

Then, let $f \in \mathcal{G} \circ \Pi$ such that $f = \sum_{A \in \mathcal{P}'} f_A \mathbf{1}_A$ for some partition $\mathcal{P}' \in \Pi$ that induces the same partition on $\mathbf{x}_1^n$ as $\mathcal{P}$. This implies that for each $A \in \mathcal{P}'$ there exists some $\bar{g}_j \in \mathcal{G}_{B_j}$ (the j is such that $A \cap \mathbf{x}_1^n = B_j$) such that

$$\frac{1}{|B_j|} \sum_{x \in B_j} |f_A(x) - \bar{g}_j(x)|^p \leq \varepsilon^p.$$

For $\bar{f} = \sum_{A \in \mathcal{P}'} \bar{g}_{A \cap \mathbf{x}_1^n} \mathbf{1}_{A \cap \mathbf{x}_1^n}$,

$$\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - \bar{f}(x_i)|^p = \frac{1}{n} \sum_j \sum_{x \in B_j} |f(x) - \bar{f}(x)|^p < \frac{1}{n} \sum_j |B_j| \varepsilon^p = \varepsilon^p.$$

This means that for each $\mathcal{P} \in \Pi$, there exists an ε cover that is no larger than $N^{M(\Pi)}$. Furthermore, by definition there are at most $\Delta(\mathbf{x}_1^n, \Pi)$ distinct partitions on $\mathbf{x}_1^n$ by members of Π. This completes the bound. ■

We are now ready to prove Theorem 1.

Proof of Theorem 1. We already established that by Theorem 2, we only need to show (7) and (8).

Lets first prove (7). Using standard analysis tools, it can be shown that m can be approximated arbitrarily closely in L_2 by functions in C_c^∞ (compactly supported) on $\mathbb{R}^d$. Thus, we can focus only on $m \in C_c^\infty$. By the assumption that $\beta_n \to \infty$, we can further restrict ourselves to $\|m\|_\infty \leq \beta_n$. Now, take $\varepsilon > 0$ and $\delta \in (0, 1)$. For some $S \subseteq \mathbb{R}^d$, conditional on the data, define $f_S \in T_{\beta_n}(\mathcal{G}_c \circ \mathcal{P}_n)$ as

$$f_S = \sum_{A \in \mathcal{P}_n} m(z_A) \mathbf{1}_{A \cap S}$$

for some $z_A \in A$ such that $z_A \in A \cap S$ if $A \cap S \neq \emptyset$. Set $\gamma > 0$ such that $|m(x) - m(z)| < \varepsilon$ for all $\|x - z\| < \gamma$. Then, for $z \in S$,

$$|f_S(z) - m(z)|^2 < \varepsilon^2 \mathbf{1}(\text{diam}(A(z) \cap S) < \gamma) + 4\|m\|_\infty^2 \mathbf{1}(\text{diam}(A(z) \cap S) \geq \gamma).$$

Then,

$$\inf_{f \in T_{\beta_n}(\mathcal{G}_c \circ \mathcal{P}_n)} \int (f(x) - m(x))^2 \mu(\mathrm{d}x)$$

$$
\begin{aligned}
&\leq \inf_{S:\mu(S)\geq 1-\delta} \int (f_S(x) - m(x))^2 \mu(\mathrm{d}x) \\
&\leq \inf_{S:\mu(S)\geq 1-\delta} \int_S |f_S(x) - m(x)|^2 \mu(\mathrm{d}x) + 4\|m\|_\infty^2 \mu(\mathbb{R}^d \backslash S) \\
&\leq \inf_{S:\mu(S)\geq 1-\delta} \int_S |f_S(x) - m(x)|^2 \mu(\mathrm{d}x) + 4\|m\|_\infty^2 \delta \\
&\leq \inf_{S:\mu(S)\geq 1-\delta} \int_S \Big(\varepsilon^2 \mathbf{1}(\mathrm{diam}(A(z)\cap S) < \gamma) + 4\|m\|_\infty^2 \mathbf{1}(\mathrm{diam}(A(z)\cap S)\geq \gamma)\Big)\mu(\mathrm{d}x) + 4\|m\|_\infty^2 \delta \\
&\leq \varepsilon^2 + 4\|m\|_\infty^2 \inf_{S:\mu(S)\geq 1-\delta} \mu(\{x\in\mathbb{R}^d : \mathrm{diam}(A(x)\cap S)\geq \gamma\}) + 4\|m\|_\infty^2 \delta \\
&\to \varepsilon^2 + 4\|m\|_\infty^2 \delta \to 0
\end{aligned}
$$

by appropriate choice of ε and δ. The last line follows by Assumption (6).

Now, we prove (3). Since $T_{\beta_n}(\mathcal{G}_c \circ \mathcal{P}_n) \subseteq T_{\beta_n}(\mathcal{G}_c \circ \Pi_n)$, it is sufficient to prove (3) for $T_{\beta_n}(\mathcal{G}_c \circ \Pi_n)$. Using Lemma 1 from Chapter 5, we can establish that

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{f\in T_{\beta_n}(\mathcal{G}_c\circ\Pi_n)} \left|\frac{1}{n}\sum |f(\mathbf{X}_i) - Y_{i,L}|^2 - \mathbb{E}[|f(\mathbf{X}) - Y_L|^2]\right| > \varepsilon\right) \\
&\leq 8\mathbb{E}\left[N\left(\frac{\varepsilon}{32\beta_n}, T_{\beta_n}(\mathcal{G}_c\circ\Pi_n), L_p(\mathbf{X}_1^n)\right)\right] \exp\left(-\frac{n\varepsilon^2}{128(4\beta_n^2)^2}\right).
\end{aligned}
$$

Then, using Lemma 1 and Theorem 12 from Chapter 5,

$$
\begin{aligned}
N\left(\frac{\varepsilon}{32\beta_n}, T_{\beta_n}(\mathcal{G}_c\circ\Pi_n), L_p(\mathbf{X}_1^n)\right) &\leq \Delta(\Pi_n) \sup_{\mathbf{z}_1^m\in\mathbf{X}_1^n, m\leq n} N\left(\frac{\varepsilon}{32\beta_n}, T_{\beta_n}(\mathcal{G}_c), L_p(\mathbf{z}_1^m)\right). \\
&\leq \Delta(\Pi_n)\left(\frac{333e\beta_n^2}{\varepsilon}\right)^{2M(\Pi)},
\end{aligned}
$$

since $V(T_{\beta_n}(\mathcal{G}_c)) \leq 1$. Thus,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{f\in T_{\beta_n}(\mathcal{G}_c\circ\Pi_n)} \left|\frac{1}{n}\sum |f(\mathbf{X}_i) - Y_{i,L}|^2 - \mathbb{E}[|f(\mathbf{X}) - Y_L|^2]\right| > \varepsilon\right) \\
&\leq 8\Delta(\Pi_n)\left(\frac{333e\beta_n^2}{\varepsilon}\right)^{2M(\Pi)} \exp\left(-\frac{n\varepsilon^2}{2048\beta_n^4}\right). \\
&\leq 8\exp\left(\log(\Delta(\Pi_n)) + 2M(\Pi_n)\log\frac{333e\beta_n^2}{\varepsilon} - \frac{n\varepsilon^2}{2048\beta_n^4}\right) \\
&= 8\exp\left(-\frac{n}{\beta_n^4}\left(\frac{\varepsilon^2}{2048} - \frac{\log(\Delta(\Pi_n))\beta_n^4}{n} - \frac{2M(\Pi_n)\beta_n^4\log\frac{333e\beta_n^2}{\varepsilon}}{n}\right)\right).
\end{aligned}
$$

The conclusion follows by the assumptions and the Borel-Cantelli lemma. ■

This theorem implies that if $\mathcal{P}_n$ is a finite data-dependent partition and $\hat{m}_n$ is the partitioning estimate (without truncation), and the following conditions are satisfied:

$$\exists k_n \in \mathbb{Z}_+ \; s.t. \; \forall A \in \mathcal{P}_n, \quad \mu_n(A) \geq k_n \frac{\log n}{n}$$
$$\lim_{n\to\infty} \operatorname{diam}(A_n(X)) \overset{p}{=} 0 \qquad \text{as } k_n \to \infty.$$

Then, $\hat{m}_n$ is weakly universally consistent. A complete proof of this implication can be found in Breiman et al. (1984).

Definition 3 (Weak universal consistency)**.** *A sequence of regression function estimates m_n is weakly universally consistent if*

$$\lim_{n\to\infty} \mathbb{E}\left[\int (m_n(x) - m(x))^2 \mu(\mathrm{d}x)\right] = 0$$

for all joint distributions of $(\mathbf{X}, Y)$ with $E[Y^2] < \infty$.

Lets now look a two concrete examples of how the consistency result can be applied.

2 Cubic partitioning estimators

Lets start with the partitioning scheme that generates partitions that are hyper-cubes, defined as $[L_n, U_n)^d$ that are then split into equidistant hyper-cubic cells.

$$\mathcal{P}_k = \{\mathbb{R}^d \backslash [L_n, U_n)^d\} \cup$$
$$\{[L_n + i_1 h_k, L_n + (i_1 + 1)h_k) \times \ldots \times [L_n + i_d h_k, L_n + (i_d + 1)h_k) : i_1, \ldots, i_d \in \{0, \ldots, k-1\}\}$$

where $h_k = (U_n - L_n)/k$ is the grid size. The partitioning scheme can be made data-dependent in the choice of k in some range $[k_{\min}, k_{\max}]$ that depends only on the sample size. Observe that this is a simplified version of CART, or more generally, decision tree estimators. Using Theorem 1, the following consistency theorem can be established for this partitioning scheme.

Theorem 3

If $L_n \to -\infty$, $U_n \to \infty$, $(U_n - L_n)/k_{\min} \to 0$,

$$\frac{(k_{\max}^d + \log n)\beta_n^4 \log(\beta_n)}{n} \to 0,$$

with $\beta_n \to \infty$, $\beta^4/n^{1-\delta} \to 0$, for some $\delta > 0$. Then, with the cubic partitioning $\mathcal{P}_n = \mathcal{P}_k$ for $k_{\min} \leq k \leq k_{\max}$, m_n is strongly universally consistent.

Definition 4 (Strong universal consistency)**.** *A sequence of regression function estimates m_n is strongly universally consistent if*

$$\lim_{n\to\infty} \int (m_n(x) - m(x))^2 \mu(\mathrm{d}x) = 0$$

with probability 1 for all joint distributions of $(\mathbf{X}, Y)$ with $E[Y^2] < \infty$.

An important observation is that Theorem 3 allows for any data-dependent selection of k. This means that standard approaches like sample-splitting or cross-validation are valid methods for selecting k that directly give a strongly consistent estimator. If we place restrictions on how k is chosen (i.e., chosen in a specific meaningful manner), we can improve the assumptions by up to log factors. It turns out that the strong consistency result of Theorem 3 also holds for the non-truncated partitioning estimator, but the proof of that result is significantly more involved.

Proof of Theorem 3. We need to verify the conditions of Theorem 1 for

$$\Pi_n = \{\mathcal{P}_k\}_{k_{\min} \le k \le k_{\max}}.$$

We can first directly compute

$$M(\Pi_n) = (k_{\max})^d + 1.$$

For $\Delta(\Pi_n)$, first condition on some $\mathbf{x}_1^n \in \mathbb{R}^d$. The partition induced by some $\mathcal{P}_k$ is uniquely determined by a vector $a_k = (a_{1,k}, \ldots, a_{n,k})$, where

$$a_{l,k} = \begin{cases} 0 & \text{if } x_l \in \mathbb{R}^d \backslash [L_n, U_n)^d \\ (i_1, \ldots, i_d) & \text{if } x_l \in [L_n + i_1 h_k, L_n + (i_1 + h_k)) \times \ldots [L_n + i_d h_k, L_n + (i_d + h_k)). \end{cases}$$

By construction, if $k_1 < k_2$, $h_{k_1} > h_{k_2}$. Thus, the corresponding allocations a_{l,k_1} and a_{l,k_2} are such that $i_1 \le j_1, \ldots i_d \le j_d$. Thus, if k varies from 1 to $k_{\max}$, there are at most $k_{\max}^d$ changes in the components of a_k. Thus, $\Delta(\mathbf{x}_1^n, \Pi_n) \le n k_{\max}^d$. These bounds of $\Delta(\Pi_n)$, $M(\Pi_n)$ can be plugged in to see that the first 4 conditions of Theorem 1 are satisfied. For the final condition, let $\gamma > 0$ and $\delta \in (0, 1)$. By the assumptions of Theorem 3, for sufficiently large n,

$$\mu([L_n, U_n)^d) \ge 1 - \delta \quad \text{and} \quad d\frac{U_n - L_n}{k_{\min}} \le \gamma.$$

Then,

$$\begin{aligned} \inf_{S : S \subseteq \mathbb{R}^d, \mu(S) \ge 1-\delta} \mu(\{x : \operatorname{diam}(A_n(x) \cap S) > \gamma\}) &\le \mu(\{x : \operatorname{diam}(A_n(x) \cap [L_n, U_n)^d) > \gamma\}) \\ &\le \mu\left(\{x : d\frac{U_n - L_n}{k_{\min}}\} > \gamma\right) \to 0 \text{ a.s.} \end{aligned}$$

for sufficiently large n. This verifies all the conditions of Theorem 1 and so the result applies to the cubic partitioning estimator. ∎

3 Nearest neighbour clustering

We can now look at an adaptive version of the kNN algorithm, where now we choose the cluster centers based on the observed data. The algorithm works as follows. There exists a function C that clusters the observations into one of k clusters with the distance metric:

$$\|x - C(x)\| = \min_{c_j : j = 1, \ldots, k} \|x - c_j\|$$

where $c_j : j = 1, \ldots, k$ are the centers of the clusters. For the nearest neighbour algorithm, ties are typically broken by choosing the center that has the minimal corresponding index j. In some cases the collection of cluster centers is known apriori. More often, however, we would like to learn the cluster centers and only pre-decide the number of clusters (possibly in a data-dependent way). In this case, we will use the squared-loss metric. Suppose $\mathbb{E}[\|X\|^2] < \infty$. The risk of the kNN algorithm is

$$R(C) = \mathbb{E}[\|X - C(x)\|^2]$$

and the empirical risk is

$$R_n(C) = \frac{1}{n}\sum_{i=1}^{n} \|X_i - C(X_i)\|^2 = \frac{1}{n}\sum_{i=1}^{n} \min_j \|X_i - c_j\|^2.$$

In order to control the complexity when the number of clusters is data-dependent, that is $k = k_n$, we will control

$$R_n(C) = \min_{|C| \le k_n} R_n(C).$$

It can be shown that this clustering scheme will exist and is obtainable for some collection of cluster centers that belong to a compact set. The consistency of this k-NN algorithm is given by the following result.

Theorem 4 (k-NN strong consistency)
Suppose $\beta_n, k_n \to \infty$ as $n \to \infty$ and

$$\frac{k_n^2 \beta_n^4 \log n}{n} \to 0, \qquad \frac{\beta_n^4}{n^{1-\delta}} \to 0$$

as $n \to \infty$ for some $\delta > 0$. Suppose m_n is constructed using the partition defined by the k-NN algorithm. Then, m_n is strongly consistent for every joint distribution of $(\mathbf{X}, Y)$ with $\mathbb{E}[\|\mathbf{X}\|^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$.

Proof of Theorem 4. For the sake of time, we will only outline the proof. Some parts of the proof will be explored in the exercises. Once again, the proof requires verifying the conditions of Theorem 1 and then invoking its conclusion. $\Delta(\Pi_{k_n})$ can be bounded by using the VC bounding arguments from Chapter 5. It should be clear that from definition, $M(\Pi_n) = k_n$. Then, it remains to verify (6). This condition requires a more careful study and involves first showing that $R_n(C_n) \to 0$ as $n \to \infty$ and the showing that diameter of any single cluster region converges to 0 as $n \to \infty$. ■

References

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. routledge.