6
k-NN Estimates

6.1 Introduction

We fix $x \in \mathcal{R}^d$, and reorder the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ according to increasing values of $\|X_i - x\|$. The reordered data sequence is denoted by

$$(X_{(1,n)}(x), Y_{(1,n)}(x)), \ldots, (X_{(n,n)}(x), Y_{(n,n)}(x))$$

or by

$$(X_{(1,n)}, Y_{(1,n)}), \ldots, (X_{(n,n)}, Y_{(n,n)})$$

if no confusion is possible. $X_{(k,n)}(x)$ is called the kth nearest neighbor (k-NN) of x.

The k_n-NN regression function estimate is defined by

$$m_n(x) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(x).$$

If X_i and X_j are equidistant from x, i.e., $\|X_i - x\| = \|X_j - x\|$, then we have a tie. There are several rules for tie breaking. For example, X_i might be declared "closer" if $i < j$, i.e., the tie breaking is done by indices. For the sake of simplicity we assume that ties occur with probability 0. In principle, this is an assumption on μ, so the statements are formally not universal, but adding a component to the observation vector X we can automatically satisfy this condition as follows: Let (X, Z) be a random vector, where Z is independent of (X, Y) and uniformly distributed on $[0, 1]$. We also artificially enlarge the data set by introducing $Z_1, Z_2, \ldots, Z_n$, where the

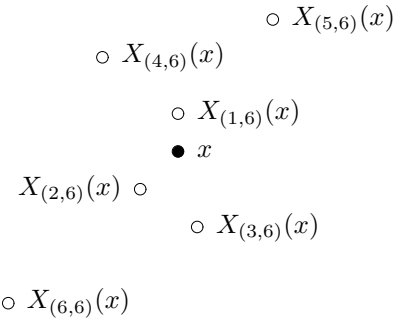

Figure 6.1. Illustration of nearest neighbors.

Z_i's are i.i.d. uniform $[0,1]$ as well. Thus, each (X_i, Z_i) is distributed as (X, Z). Then ties occur with probability 0. In the sequel we shall assume that X has such a component and, therefore, for each x the random variable $\|X - x\|^2$ is absolutely continuous, since it is a sum of two independent random variables such that one of the two is absolutely continuous.

Figures 6.2 – 6.4 show k_n-NN estimates for various choices of k_n for our simulated data introduced in Chapter 1. Figure 6.5 shows the L_2 error as a function of k_n.

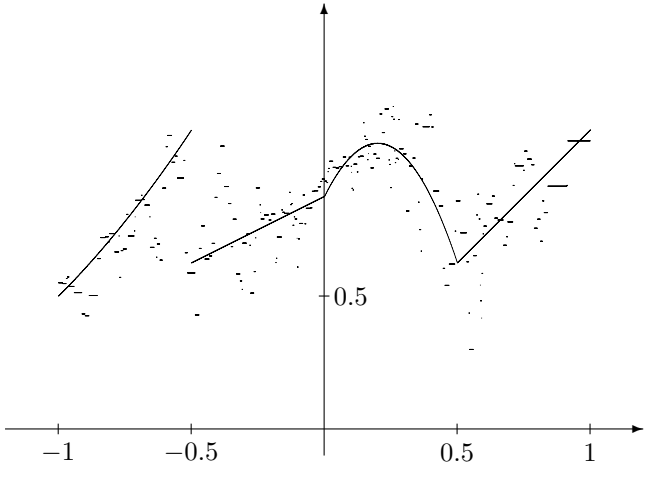

Figure 6.2. Undersmoothing: $k_n = 3$, L_2 error =0.011703.

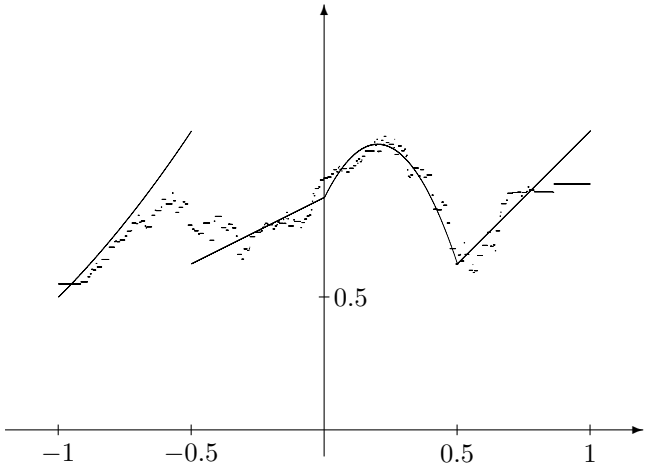

Figure 6.3. Good choice: $k_n = 12$, L_2 error =0.004247.

6.2 Consistency

In this section we use Stone's theorem (Theorem 4.1) in order to prove weak universal consistency of the k-NN estimate. The main result is the following theorem:

Theorem 6.1. *If $k_n \to \infty$, $k_n/n \to 0$, then the k_n-NN regression function estimate is weakly consistent for all distributions of (X, Y) where ties occur with probability zero and $\mathbf{E}Y^2 < \infty$.*

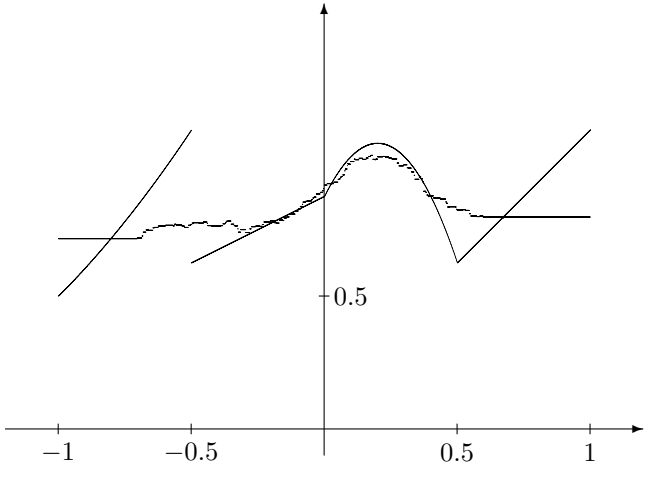

Figure 6.4. Oversmoothing: $k_n = 50$, L_2 error =0.009931.

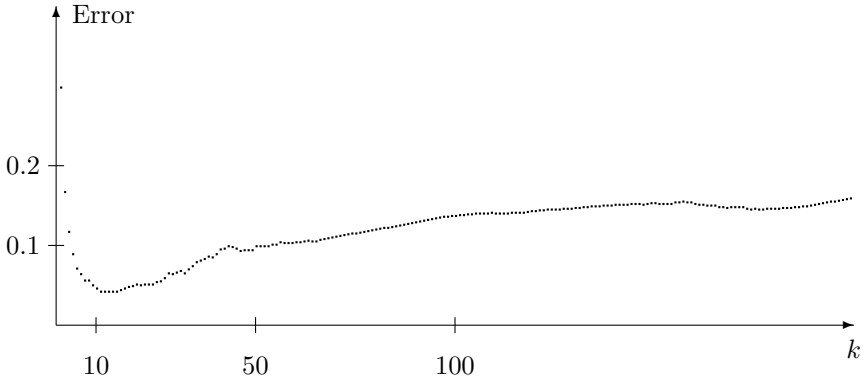

Figure 6.5. L_2 error of the k-NN estimate as a function of k.

According to Theorem 6.1 the number of nearest neighbors (k_n), over which one averages in order to estimate the regression function, should on the one hand converge to infinity but should, on the other hand, be small with respect to the sample size n. To verify the conditions of Stone's theorem we need several lemmas.

We will use Lemma 6.1 to verify condition (iii) of Stone's theorem. Denote the probability measure for X by μ, and let $S_{x,\epsilon}$ be the closed ball centered at x of radius $\epsilon > 0$. The collection of all x with $\mu(S_{x,\epsilon}) > 0$ for all $\epsilon > 0$ is called the support of X or μ. This set plays a key role because of the following property:

Lemma 6.1. *If $x \in$ support(μ) and $\lim_{n\to\infty} k_n/n = 0$, then*

$$\|X_{(k_n,n)}(x) - x\| \to 0$$

with probability one.

PROOF. Take $\epsilon > 0$. By definition, $x \in \text{support}(\mu)$ implies that $\mu(S_{x,\epsilon}) > 0$. Observe that

$$\{\|X_{(k_n,n)}(x) - x\| > \epsilon\} = \left\{ \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in S_{x,\epsilon}\}} < \frac{k_n}{n} \right\}.$$

By the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \in S_{x,\epsilon}\}} \to \mu(S_{x,\epsilon}) > 0$$

with probability one, while, by assumption,

$$\frac{k_n}{n} \to 0.$$

Therefore, $\|X_{(k_n,n)}(x) - x\| \to 0$ with probability one. □

The next two lemmas will enable us to establish condition (i) of Stone's theorem.

Lemma 6.2. *Let*

$$B_a(x') = \left\{x : \mu(S_{x,\|x-x'\|}) \le a\right\}.$$

Then, for all $x' \in \mathcal{R}^d$,

$$\mu(B_a(x')) \le \gamma_d a,$$

where γ_d *depends on the dimension* d *only.*

PROOF. Let $C_j \subset \mathcal{R}^d$ be a cone of angle $\pi/3$ and centered at 0. It is a property of cones that if $u, u' \in C_j$ and $\|u\| < \|u'\|$, then $\|u - u'\| < \|u'\|$ (cf. Figure 6.6). Let $C_1, \ldots, C_{\gamma_d}$ be a collection of such cones with different central directions such that their union covers $\mathcal{R}^d$:

$$\bigcup_{j=1}^{\gamma_d} C_j = \mathcal{R}^d.$$

Then

$$\mu(B_a(x')) \le \sum_{i=1}^{\gamma_d} \mu(\{x' + C_i\} \cap B_a(x')).$$

Let $x^* \in \{x' + C_i\} \cap B_a(x')$. Then, by the property of cones mentioned above, we have

$$\mu(\{x' + C_i\} \cap S_{x',\|x'-x^*\|} \cap B_a(x')) \le \mu(S_{x^*,\|x'-x^*\|}) \le a,$$

where we use the fact that $x^* \in B_a(x')$. Since x^* is arbitrary,

$$\mu(\{x' + C_i\} \cap B_a(x')) \le a,$$

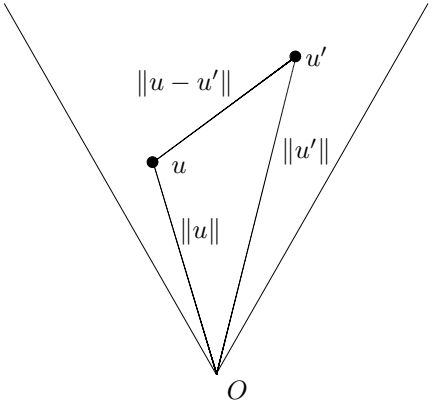

Figure 6.6. The cone property.

which completes the proof of the lemma. □

An immediate consequence of the lemma is that the number of points among $X_1, \ldots, X_n$, such that X is one of their k nearest neighbors, is not more than a constant times k.

Corollary 6.1. *Assume that ties occur with probability zero. Then*

$$\sum_{i=1}^{n} I_{\{X \text{ is among the } k \text{ NNs of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}\}} \le k\gamma_d$$

a.s.

PROOF. Apply Lemma 6.2 with $a = k/n$ and let μ be the empirical measure μ_n of $X_1, \ldots, X_n$, i.e., for each Borel set $A \subseteq \mathcal{R}^d$, $\mu_n(A) = (1/n)\sum_{i=1}^{n} I_{\{X_i \in A\}}$. Then

$$B_{k/n}(X) = \left\{x \,:\, \mu_n(S_{x,\|x-X\|}) \le k/n\right\}$$

and

$$\begin{aligned} & X_i \in B_{k/n}(X) \\ \Leftrightarrow \quad & \mu_n(S_{X_i,\|X_i-X\|}) \le k/n \\ \Leftrightarrow \quad & X \text{ is among the } k \text{ NNs of } X_i \text{ in } \{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\} \end{aligned}$$

a.s., where for the second $\Leftrightarrow$ we applied the condition that ties occur with probability zero. This, together with Lemma 6.2, yields

$$\begin{aligned} & \sum_{i=1}^{n} I_{\{X \text{ is among the } k \text{ NNs of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}\}} \\ = \quad & \sum_{i=1}^{n} I_{\{X_i \in B_{k/n}(X)\}} \\ = \quad & n \cdot \mu_n(B_{k/n}(X)) \\ \le \quad & k\gamma_d \end{aligned}$$

a.s. □

Lemma 6.3. *Assume that ties occur with probability zero. Then for any integrable function f, any n, and any $k \le n$,*

$$\sum_{i=1}^{k} \mathbf{E}\left\{|f(X_{(i,n)}(X))|\right\} \le k\gamma_d \mathbf{E}\{|f(X)|\},$$

where γ_d depends upon the dimension only.

PROOF. If f is a nonnegative function,

$$\begin{aligned}
&\sum_{i=1}^{k} \mathbf{E}\left\{f(X_{(i,n)}(X))\right\} \\
= \quad & \mathbf{E}\left\{\sum_{i=1}^{n} I_{\{X_i \text{ is among the k NNs of } X \text{ in } \{X_1,\ldots,X_n\}\}} f(X_i)\right\} \\
= \quad & \mathbf{E}\left\{f(X)\sum_{i=1}^{n} I_{\{X \text{ is among the k NNs of } X_i \text{ in } \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}\}}\right\} \\
& \text{(by exchanging } X \text{ and } X_i\text{)} \\
\leq \quad & \mathbf{E}\{f(X)k\gamma_d\},
\end{aligned}$$

by Corollary 6.1. This concludes the proof of the lemma. □

PROOF OF THEOREM 6.1. We proceed by checking the conditions of Stone's weak convergence theorem (Theorem 4.1) under the condition that ties occur with probability zero. The weight $W_{n,i}(X)$ in Theorem 4.1 equals $1/k_n$ if X_i is among the k_n nearest neighbors of X, and equals 0 otherwise, thus the weights are probability weights, and (ii) and (iv) are automatically satisfied. Condition (v) is obvious since $k_n \to \infty$. For condition (iii) observe that, for each $\epsilon > 0$,

$$\begin{aligned}
& \mathbf{E}\left\{\sum_{i=1}^{n} W_{n,i}(X) I_{\{\|X_i - X\| > \epsilon\}}\right\} \\
= \quad & \int \mathbf{E}\left\{\sum_{i=1}^{n} W_{n,i}(x) I_{\{\|X_i - x\| > \epsilon\}}\right\} \mu(dx) \\
= \quad & \int \mathbf{E}\left\{\frac{1}{k_n}\sum_{i=1}^{k_n} I_{\{\|X_{(i,n)}(x) - x\| > \epsilon\}}\right\} \mu(dx) \to 0
\end{aligned}$$

holds whenever

$$\int \mathbf{P}\left\{\|X_{(k_n,n)}(x) - x\| > \epsilon\right\} \mu(dx) \to 0, \tag{6.1}$$

where $X_{(k_n,n)}(x)$ denotes the k_nth nearest neighbor of x among $X_1, \ldots, X_n$. For $x \in \mathit{support}(\mu)$, $k_n/n \to 0$, together with Lemma 6.1, implies

$$\mathbf{P}\left\{\|X_{(k_n,n)}(x) - x\| > \epsilon\right\} \to 0 \quad (n \to \infty).$$

This together with the dominated convergence theorem implies (6.1). Finally, we consider condition (i). It suffices to show that for any nonnegative measurable function f with $\mathbf{E}\{f(X)\} < \infty$, and any n,

$$\mathbf{E}\left\{\sum_{i=1}^{n} \frac{1}{k_n} I_{\{X_i \text{ is among the } k_n \text{ NNs of } X\}} f(X_i)\right\} \leq c \cdot \mathbf{E}\left\{f(X)\right\}$$

for some constant c. But we have shown in Lemma 6.3 that this inequality always holds with $c = \gamma_d$. Thus, condition (i) is verified. □

6.3 Rate of Convergence

In this section we bound the rate of convergence of $\mathbf{E}\|m_n - m\|^2$ for a k_n-nearest neighbor estimate.

Theorem 6.2. *Assume that X is bounded,*

$$\sigma^2(x) = \mathbf{Var}(Y|X = x) \le \sigma^2 \quad (x \in \mathcal{R}^d)$$

and

$$|m(x) - m(z)| \le C\|x - z\| \quad (x, z \in \mathcal{R}^d).$$

Assume that $d \ge 3$. Let m_n be the k_n-NN estimate. Then

$$\mathbf{E}\|m_n - m\|^2 \le \frac{\sigma^2}{k_n} + c_1 \cdot C^2 \left(\frac{k_n}{n}\right)^{2/d},$$

thus for $k_n = c' \left(\sigma^2/C^2\right)^{d/(2+d)} n^{\frac{2}{d+2}}$,

$$\mathbf{E}\|m_n - m\|^2 \le c'' \sigma^{\frac{4}{d+2}} C^{\frac{2d}{2+d}} n^{-\frac{2}{d+2}}.$$

For the proof of Theorem 6.2 we need the rate of convergence of nearest neighbor distances.

Lemma 6.4. *Assume that X is bounded. If $d \ge 3$, then*

$$\mathbf{E}\{\|X_{(1,n)}(X) - X\|^2\} \le \frac{\tilde{c}}{n^{2/d}}.$$

PROOF. For fixed $\epsilon > 0$,

$$\mathbf{P}\{\|X_{(1,n)}(X) - X\| > \epsilon\} = \mathbf{E}\{(1 - \mu(S_{X,\epsilon}))^n\}.$$

Let $A_1, \ldots, A_{N(\epsilon)}$ be a cubic partition of the bounded support of μ such that the A_j's have diameter ϵ and

$$N(\epsilon) \le \frac{c}{\epsilon^d}.$$

If $x \in A_j$, then $A_j \subset S_{x,\epsilon}$, therefore

$$\begin{aligned} \mathbf{E}\{(1 - \mu(S_{X,\epsilon}))^n\} &= \sum_{j=1}^{N(\epsilon)} \int_{A_j} (1 - \mu(S_{x,\epsilon}))^n \mu(dx) \\ &\le \sum_{j=1}^{N(\epsilon)} \int_{A_j} (1 - \mu(A_j))^n \mu(dx) \end{aligned}$$

$$= \sum_{j=1}^{N(\epsilon)} \mu(A_j)(1-\mu(A_j))^n.$$

Obviously,

$$\begin{aligned}\sum_{j=1}^{N(\epsilon)} \mu(A_j)(1-\mu(A_j))^n &\le \sum_{j=1}^{N(\epsilon)} \max_z z(1-z)^n \\ &\le \sum_{j=1}^{N(\epsilon)} \max_z z e^{-nz} \\ &= \frac{e^{-1}N(\epsilon)}{n}.\end{aligned}$$

If L stands for the diameter of the support of μ, then

$$\begin{aligned}\mathbf{E}\{\|X_{(1,n)}(X)-X\|^2\} &= \int_0^\infty \mathbf{P}\{\|X_{(1,n)}(X)-X\|^2 > \epsilon\}\, d\epsilon \\ &= \int_0^{L^2} \mathbf{P}\{\|X_{(1,n)}(X)-X\| > \sqrt{\epsilon}\}\, d\epsilon \\ &\le \int_0^{L^2} \min\left\{1, \frac{e^{-1}N(\sqrt{\epsilon})}{n}\right\} d\epsilon \\ &\le \int_0^{L^2} \min\left\{1, \frac{c}{en}\epsilon^{-d/2}\right\} d\epsilon \\ &= \int_0^{(c/(en))^{2/d}} 1\, d\epsilon + \frac{c}{en}\int_{(c/(en))^{2/d}}^{L^2} \epsilon^{-d/2} d\epsilon \\ &\le \frac{\tilde{c}}{n^{2/d}}\end{aligned}$$

for $d \ge 3$. □

Proof of Theorem 6.2. We have the decomposition

$$\begin{aligned}\mathbf{E}\{(m_n(x)-m(x))^2\} &= \mathbf{E}\{(m_n(x)-\mathbf{E}\{m_n(x)|X_1,\ldots,X_n\})^2\} \\ &\quad + \mathbf{E}\{(\mathbf{E}\{m_n(x)|X_1,\ldots,X_n\}-m(x))^2\} \\ &= I_1(x)+I_2(x).\end{aligned}$$

The first term is easier:

$$\begin{aligned}I_1(x) &= \mathbf{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\left(Y_{(i,n)}(x)-m(X_{(i,n)}(x))\right)\right)^2\right\} \\ &= \mathbf{E}\left\{\frac{1}{k_n^2}\sum_{i=1}^{k_n}\sigma^2(X_{(i,n)}(x))\right\}\end{aligned}$$

$$\leq \frac{\sigma^2}{k_n}.$$

For the second term

$$\begin{aligned} I_2(x) &= \mathbf{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(m(X_{(i,n)}(x)) - m(x))\right)^2\right\} \\ &\leq \mathbf{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}|m(X_{(i,n)}(x)) - m(x)|\right)^2\right\} \\ &\leq \mathbf{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}C\|X_{(i,n)}(x) - x\|\right)^2\right\}. \end{aligned}$$

Put $N = k_n\lfloor\frac{n}{k_n}\rfloor$. Split the data $X_1, \ldots, X_n$ into $k_n + 1$ segments such that the first k_n segments have length $\lfloor\frac{n}{k_n}\rfloor$, and let $\tilde{X}_j^x$ be the first nearest neighbor of x from the jth segment. Then $\tilde{X}_1^x$, $\ldots$, $\tilde{X}_{k_n}^x$ are k_n different elements of $\{X_1, \ldots, X_n\}$, which implies

$$\sum_{i=1}^{k_n}\|X_{(i,n)}(x) - x\| \leq \sum_{j=1}^{k_n}\|\tilde{X}_j^x - x\|,$$

therefore, by Jensen's inequality,

$$\begin{aligned} I_2(x) &\leq C^2\mathbf{E}\left\{\left(\frac{1}{k_n}\sum_{j=1}^{k_n}\|\tilde{X}_j^x - x\|\right)^2\right\} \\ &\leq C^2\frac{1}{k_n}\sum_{j=1}^{k_n}\mathbf{E}\left\{\|\tilde{X}_j^x - x\|^2\right\} \\ &= C^2\mathbf{E}\left\{\|\tilde{X}_1^x - x\|^2\right\} \\ &= C^2\mathbf{E}\left\{\|X_{(1,\lfloor\frac{n}{k_n}\rfloor)}(x) - x\|^2\right\}. \end{aligned}$$

Thus, by Lemma 6.4,

$$\begin{aligned} \frac{1}{C^2}\left\lfloor\frac{n}{k_n}\right\rfloor^{2/d}\int I_2(x)\mu(dx) &\leq \left\lfloor\frac{n}{k_n}\right\rfloor^{2/d}\mathbf{E}\left\{\|X_{(1,\lfloor\frac{n}{k_n}\rfloor)}(X) - X\|^2\right\} \\ &\leq const. \end{aligned}$$

□

For $d \leq 2$ the rate of convergence of Theorem 6.2 holds under additional conditions on μ (cf. Problem 6.7).

According to Theorem 6.2, the nearest neighbor estimate is of optimum rate for the class $\mathcal{D}^{(1,C)}$ (cf. Definition 3.2 and Theorem 3.2). In Theorem

6.2 the only condition on X for $d \geq 3$ is that it has compact support, there is no density assumption.

Similarly to the partitioning estimate, the nearest neighbor estimate cannot "track" the derivative of a differentiable regression function. In the pointwise theory the nearest neighbor regression estimate has the optimum rate of convergence for the class $\mathcal{D}^{(2,C)}$ (cf. Härdle (1990)). Unfortunately, in the L_2 theory, this is not the case.

In order to show this consider the following example:
- X is uniform on $[0,1]$;
- $m(x) = x$; and
- $Y = X + N$, where N is standard normal and is independent of X.

This example belongs to $\mathcal{D}^{(p,C)}$ for any $p \geq 1$. In Problem 6.2 we will see that, for $k_n/n \to 0$ and $k_n \to \infty$,

$$\mathbf{E} \int_0^1 (m_n(x) - m(x))^2 \mu(dx) \geq \frac{1}{k_n} + \frac{1}{24} \left(\frac{k_n}{n+1} \right)^3, \tag{6.2}$$

where the lower bound is minimized by $k_n = cn^{3/4}$, and thus

$$\mathbf{E} \int_0^1 (m_n(x) - m(x))^2 \mu(dx) \geq c' n^{-\frac{3}{4}},$$

therefore the nearest neighbor estimate is not optimal for $\mathcal{D}^{(2,C)}$.

The main point of this example is that because of the end points of the uniform density the squared bias is of order $\left(\frac{k_n}{n}\right)^3$ and not $\left(\frac{k_n}{n}\right)^4$. From this one may conjecture that the nearest neighbor regression estimate is optimal for the class $\mathcal{D}^{(1.5,C)}$.

6.4 Bibliographic Notes

The consistency of the k_n-nearest neighbor classification, and the corresponding regression and density estimation has been studied by many researchers. See Beck (1979), Bhattacharya and Mack (1987), Bickel and Breiman (1983), Cheng (1995), Collomb (1979; 1980; 1981), Cover (1968a), Cover and Hart (1967), Devroye (1978a; 1981; 1982b), Devroye and Györfi (1985), Devroye et al. (1994), Fix and Hodges (1951; 1952), Guerre (2000) Györfi and Györfi (1975), Mack (1981), Stone (1977), Stute (1984), and Zhao (1987). Theorem 6.1 is due to Stone (1977). Various versions of Lemma 6.2 appeared in Fritz (1974), Stone (1977), Devroye and Györfi (1985). Lemma 6.4 is a special case of the result of Kulkarni and Posner (1995).

Problems and Exercises

PROBLEM 6.1. Prove that for $d \leq 2$ Lemma 6.4 is not distribution-free, i.e., construct a distribution of X for which Lemma 6.4 does not hold.

HINT: Put $d = 1$ and assume a density $f(x) = 3x^2$, then $F(x) = x^3$ and

$$\begin{aligned}\mathbf{E}\{\|X_{(1,n)}(X) - X\|^2\} &\geq \int_0^{1/4}\int_0^{\sqrt{\epsilon}} (1 - [F(x+\sqrt{\epsilon}) - F(x-\sqrt{\epsilon})])^n f(x)\,dx\,d\epsilon \\ &\geq \frac{C}{n^{5/3}}.\end{aligned}$$

PROBLEM 6.2. Prove (6.2).

HINT:

Step (a).

$$\begin{aligned}&\mathbf{E}(m_n(x) - m(x))^2 \\ \geq\ &\mathbf{E}(m_n(x) - \mathbf{E}\{m_n(x)|X_1, \ldots X_n\})^2 + (\mathbf{E}\{m_n(x)\} - m(x))^2.\end{aligned}$$

Step (b).

$$\mathbf{E}\{(m_n(x) - \mathbf{E}\{m_n(x)|X_1, \ldots X_n\})^2\} = \frac{1}{k_n}.$$

Step (c). Observe that the function

$$\mathbf{E}\{m_n(x)|X_1, \ldots, X_n\} = \frac{1}{k_n}\sum_{i=1}^{k_n} X_{(i,n)}(x)$$

is a monotone increasing function of x, therefore

$$\mathbf{E}\{m_n(x)|X_1, \ldots, X_n\} \geq \frac{1}{k_n}\sum_{i=1}^{k_n} X_{(i,n)}(0).$$

Let $X_1^*, \ldots, X_n^*$ be the ordered sample of $X_1, \ldots, X_n$, then $X_{(i,n)}(0) = X_i^*$, and so

$$\mathbf{E}\{m_n(x)\} \geq \mathbf{E}\left\{\frac{1}{k_n}\sum_{i=1}^{k_n} X_i^*\right\} = \alpha_{k_n}.$$

Thus

$$\int_0^1 (\mathbf{E}\{m_n(x)\} - m(x))^2 \mu(dx) \geq \frac{\alpha_{k_n}^3}{3}.$$

Step (d).

$$\alpha_{k_n} = \frac{1}{2}\frac{k_n}{n+1}.$$

PROBLEM 6.3. Prove that for fixed k the k-NN regression estimate is weakly consistent for noiseless observations.

HINT: See Problem 4.5.

PROBLEM 6.4. Let $m_n(x)$ be the k-NN regression estimate. Prove that, for fixed k,

$$\lim_{n\to\infty} \mathbf{E} \int (m_n(x) - m(x))^2 \mu(dx) = \frac{\mathbf{E}(Y - m(X))^2}{k}$$

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

HINT: Use the decomposition

$$m_n(x) = \frac{1}{k} \sum_{i=1}^{k} m(X_{(i,n)}(x)) + \frac{1}{k} \sum_{i=1}^{k} (Y_{(i,n)}(x) - m(X_{(i,n)}(x))).$$

Handle the first term by Problem 6.3. Show that

$$\begin{aligned} \mathbf{E} \int \left(\frac{1}{k} \sum_{i=1}^{k} (Y_{(i,n)}(x) - m(X_{(i,n)}(x))) \right)^2 \mu(dx) &= \frac{1}{k^2} \sum_{i=1}^{k} \mathbf{E}\{\sigma^2(X_{(i,n)}(X))\} \\ &\to \frac{\mathbf{E}(Y - m(X))^2}{k}. \end{aligned}$$

PROBLEM 6.5. Let g_n be the k-NN classification rule for M classes:

$$g_n(x) = \arg\max_{1 \le j \le M} \sum_{i=1}^{k} I_{\{Y_{(i,n)}(x)=j\}}.$$

Show that, for $k_n \to \infty$ and $k_n/n \to 0$,

$$\lim_{n\to\infty} \mathbf{P}\{g_n(X) \neq Y\} = \mathbf{P}\{g^*(X) \neq Y\}$$

for all distributions of (X, Y), where g^* is the Bayes decision rule (Devroye, Györfi, and Lugosi (1996)).

HINT: Apply Problem 1.5 and Theorem 6.1.

PROBLEM 6.6. Let g_n be the 1-NN classification rule. Prove that

$$\lim_{n\to\infty} \mathbf{P}\{g_n(X) \neq Y\} = 1 - \sum_{j=1}^{M} \mathbf{E}\{m^{(j)}(X)^2\}$$

for all distributions of (X, Y), where $m^{(j)}(X) = \mathbf{P}\{Y = j|X\}$ (Cover and Hart (1967), Stone (1977)).

HINT:

Step (a). Show that

$$\begin{aligned} \mathbf{P}\{g_n(X) \neq Y\} &= 1 - \sum_{j=1}^{M} \mathbf{P}\{Y = j, g_n(X) = j\} \\ &= 1 - \sum_{j=1}^{M} \mathbf{E}\{m^{(j)}(X) m^{(j)}(X_{(1,n)}(X))\}. \end{aligned}$$

Step (b). Problem 6.3 implies that

$$\lim_{n\to\infty} \mathbf{E}\{(m^{(j)}(X) - m^{(j)}(X_{(1,n)}(X)))^2\} = 0.$$

PROBLEM 6.7. For $d \leq 2$ assume that there exist $\epsilon_0 > 0$, a nonnegative function g such that for all $x \in \mathcal{R}^d$, and $0 < \epsilon \leq \epsilon_0$,

$$\mu(S_{x,\epsilon}) > g(x)\epsilon^d \tag{6.3}$$

and

$$\int \frac{1}{g(x)^{2/d}} \mu(dx) < \infty.$$

Prove the rate of convergence given in Theorem 6.2.

HINT: Prove that under the conditions of the problem

$$\mathbf{E}\{\|X_{(1,n)}(X) - X\|^2\} \leq \frac{\tilde{c}}{n^{2/d}}.$$

Formula (6.3) implies that for almost all x mod μ and $\epsilon_0 < \epsilon < L$,

$$\mu(S_{x,\epsilon}) \geq \mu(S_{x,\epsilon_0}) \geq g(x)\epsilon_0^d \geq g(x) \left(\frac{\epsilon_0}{L}\right)^d \epsilon^d,$$

hence we can assume w.l.o.g. that (6.3) holds for all $0 < \epsilon < L$. In this case, we get, for fixed $L > \epsilon > 0$,

$$\begin{aligned} \mathbf{P}\{\|X_{(1,n)}(X) - X\| > \epsilon\} &= \mathbf{E}\{(1 - \mu(S_{X,\epsilon}))^n\} \\ &\leq \mathbf{E}\left\{e^{-n\mu(S_{X,\epsilon})}\right\} \\ &\leq \mathbf{E}\left\{e^{-ng(X)\epsilon^d}\right\}, \end{aligned}$$

therefore,

$$\begin{aligned} \mathbf{E}\{\|X_{(1,n)}(X) - X\|^2\} &= \int_0^{L^2} \mathbf{P}\{\|X_{(1,n)}(X) - X\| > \sqrt{\epsilon}\}\, d\epsilon \\ &\leq \int_0^{L^2} \mathbf{E}\left\{e^{-ng(X)\epsilon^{d/2}}\right\} d\epsilon \\ &\leq \int \int_0^{\infty} e^{-ng(x)\epsilon^{d/2}}\, d\epsilon\, \mu(dx) \\ &= \int \frac{1}{n^{2/d} g(x)^{2/d}} \int_0^{\infty} e^{-z^{d/2}}\, dz\, \mu(dx) \\ &= \frac{\tilde{c}}{n^{2/d}}. \end{aligned}$$