02 kNN & Kernels

We saw previously that, without any structural assumptions, regression estimation is essentially hopeless, no matter the estimator. However, there are interesting/non-trivial nonparametric assumptions under which substantial progress can be made. Here we present some relatively traditional flexible regression methods (based on local versus global modeling), and analyze their mean squared error under smoothness assumptions.

1 kNN and kernels

In previous courses you have seen k-nearest neighbor and simple kernel methods. Here we review these methods and discuss some of their properties.

The k-nearest neighbor estimator simply estimates the regression function at a point x by averaging the outcomes of the k nearest neighbors, typically defined via Euclidean distance. To define this estimator first let

$$\Big\{D_1(x), ..., D_n(x)\Big\} = \Big\{ \min_i \|X_i - x\|, ..., \max_i \|X_i - x\| \Big\}$$

denote the sorted distances $\|X_i - x\|$, $i = 1, ..., n$, and then the k-nearest neighbor estimator is

$$\widehat{\mu}(x) = \frac{1}{k} \sum_{i=1}^{n} Y_i \, \mathbb{1}\Big\{ \|X_i - x\| \leq D_k(x) \Big\}.$$

Note that $k = \sum_{i=1}^{n} \mathbb{1}\{\|X_i - x\| \leq D_k(x)\}$ by definition (we assume there are no ties for simplicity).

The k-nearest neighbor estimator is used quite often in practice perhaps due to its simplicity and interpretability. As seen below, it has some useful properties which ensure it is *universally consistent* (i.e., it has L_2 risk tending to zero for any distribution with $\mathbb{E}(Y^2)$ finite) and optimal for functions with relatively low smoothness.

Theorem 1 (Theorems 6.1 and 6.2, Györfi et al. [2002])**.** *Let $\widehat{\mu}(x)$ denote the k-nearest neighbor estimator defined above. If $k/n \to 0$ and $k \to \infty$ then*

$$\lim_{n \to \infty} \mathbb{E}\|\widehat{\mu} - \mu\| = 0$$

for any distribution $\mathbb{P}$ such that $\mathbb{E}(Y^2) < \infty$ (and X is continuously distributed). If $var(Y \mid X = x) \leq \sigma^2$ and μ is Lipschitz (i.e., $|\mu(x) - \mu(x')| \leq L\|x - x'\|$) then

$$\mathbb{E}\|\widehat{\mu} - \mu\| \leq CL \left(\frac{k}{n}\right)^{1/d} + \frac{\sigma}{\sqrt{k}}$$

whenever $d \geq 3$, for some constant C.

The Nadaraya-Watson or local-constant kernel estimator is a simple weighted average estimator, much like k-nearest neighbors, but where instead of a fixed number of observations

being used, a fixed window is used, defined by a kernel function $K : \mathbb{R}^d \to \mathbb{R}$ and a bandwidth parameter h. Common kernels include the boxcar kernel $K(u) = \mathbb{1}(\|u\| \leq 1)$, the Epanechnikov kernel $K(u) = (1 - \|u\|^2)_+$, and the Gaussian kernel $K(u) = \exp(-\|u\|^2)$. Then the kernel estimate is given by

$$\widehat{\mu}(x) = \frac{\sum_i Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_i K\left(\frac{X_i - x}{h}\right)}.$$

For example, for the boxcar kernel, this is just the average outcome among all units i with $\|X_i - x\| \leq h$, i.e., that are contained in a ball of radius h around x.

The following graphic from Hastie et al. illustrates k-nearest neighbor and kernel estimators (with Epanechnikov kernel) in a simple one-dimensional example.

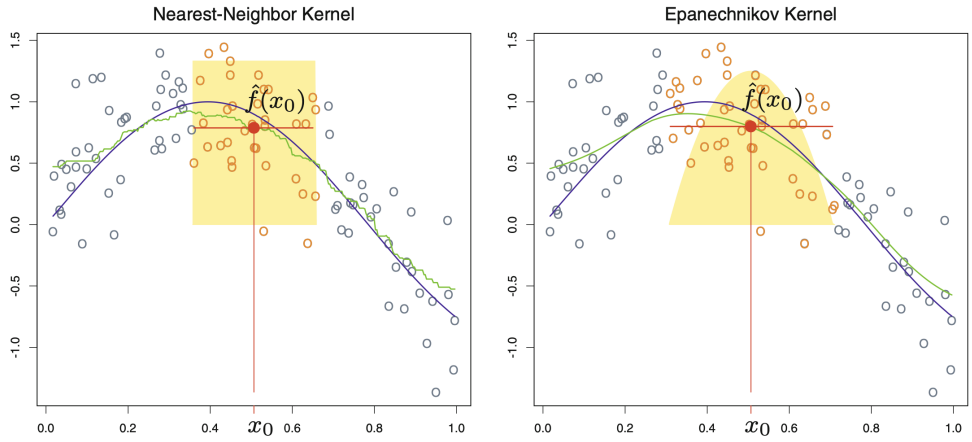

FIGURE 6.1. *In each panel* 100 *pairs* x_i, y_i *are generated at random from the blue curve with Gaussian errors:* $Y = \sin(4X) + \varepsilon$, $X \sim U[0, 1]$, $\varepsilon \sim N(0, 1/3)$. *In the left panel the green curve is the result of a* 30*-nearest-neighbor running-mean smoother. The red point is the fitted constant* $\hat{f}(x_0)$, *and the red circles indicate those observations contributing to the fit at* x_0. *The solid yellow region indicates the weights assigned to observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width* $\lambda = 0.2$.

Just like the k-nearest neighbor estimator, the kernel estimator is also simple/interpretable and used quite often in practice. It is also universally consistent under mild conditions and optimal for functions with relatively low smoothness.

Theorem 2 (Theorems 5.1 and 5.2, Györfi et al. [2002]). *Let* $\widehat{\mu}(x)$ *denote the kernel estimator defined above. If* $h \to 0$ *and* $nh^d \to \infty$ *then*

$$\lim_{n \to \infty} \mathbb{E}\|\widehat{\mu} - \mu\| = 0$$

for any distribution $\mathbb{P}$ *such that* $\mathbb{E}(Y^2) < \infty$, *provided the kernel satisfies the boxed kernel condition that*

$$b\mathbb{1}(\|u\| \leq r) \leq K(u) \leq \mathbb{1}(\|u\| \leq R)$$

for some constant $b > 0$ and radii $0 < r \leq R$. If X has compact support, $var(Y \mid X = x) \leq \sigma^2$, and μ is Lipschitz (i.e., $|\mu(x) - \mu(x')| \leq L\|x - x'\|$) and bounded ($\sup_x |\mu(x)| \leq M$), then

$$\mathbb{E}\|\widehat{\mu} - \mu\| \leq Lh + \frac{C(\sigma + M)}{\sqrt{nh^d}}$$

for some constant C, when K is the boxcar kernel.

The boxed kernel condition in the first part of the Theorem is illustrated in the following image from Györfi et al. [2002], and shows that the condition would be satisfied for any kernel that is bounded away from zero at the origin and has compact support (note this includes boxcar and Epanechnikov kernels, but not Gaussian kernels).

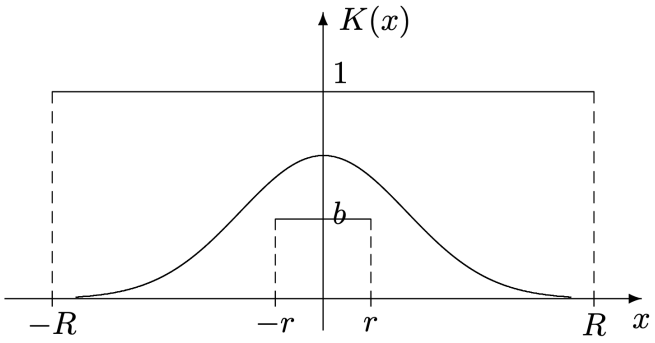

Figure 5.7. Boxed kernel.

2 Linear smoothers and Stone's theorem

Note that both the k-nearest neighbor and kernel estimators are weighted averages of the form

$$\widehat{\mu}(x) = \sum_{i=1}^{n} w_i(x; X^n) Y_i$$

where the weights $w_i(x; X^n)$ can depend on all the features $X^n = (X_1, ..., X_n)$ but not the outcomes. For k-nearest neighbors the weights take the form

$$w_i(x; X^n) = \frac{1}{k} \mathbb{1}\Big\{ \|X_i - x\| \leq D_k(x) \Big\}$$

while for kernel estimators the weights take the form

$$w_i(x; X^n) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_i K\left(\frac{X_i - x}{h}\right)}.$$

In fact, estimators of this form are called *linear smoothers*, and they make up a very important class of regression estimators, which appears all over nonparametrics and machine learning.

Stone [1977] gave the following useful theorem that can be used to prove universal consistency for linear smoothers (Györfi et al. [2002] detail how it can be used with k-nearest neighbors, and kernel estimators, among other methods).

Theorem 3 (Stone's Theorem (Theorem 4.1, Györfi et al. [2002])). *Let $\widehat{\mu}(x) = \sum_i w_i(x; X^n) Y_i$ be a linear smoother. If the weights satisfy, for an independent draw X, that:*

1. $\mathbb{E} \sum_i |w_i(X; X^n)| f(X_i) \leq C \mathbb{E}\{f(X)\}$ *for all* $f \geq 0$ *with* $\mathbb{E}\{f(X)\} < \infty$ *and some* $C > 0$,
2. $\sum_i |w_i(X; X^n)| \leq C$ *with probability one for some* $C \geq 1$,
3. $\sum_i w_i(X; X^n) \xrightarrow{p} 1$,
4. $\mathbb{E} \sum_i |w_i(X; X^n)| \mathbb{1}(\|X_i - X\| > a) \to 0$ *as* $n \to \infty$ *for all* $a > 0$, *and*
5. $\mathbb{E} \sum_i w_i(X; X^n)^2 \to 0$ *as* $n \to \infty$.

then

$$\lim_{n \to \infty} \mathbb{E}\|\widehat{\mu} - \mu\| = 0$$

for all distributions with $\mathbb{E}(Y^2) < \infty$ *(i.e., $\widehat{\mu}$ is universally consistent).*

Condition 1 says the mean estimate can't be too much larger than the mean regression function; without it, the result will still hold if the regression function is uniformly continuous and the conditional variance is bounded. Conditions 2 and 3 say that the sum of the weights should be bounded and asymptotically one. Condition 4 says the weights should be asymptotically localized (i.e., only influenced by points near X). Condition 5 says that all the weights should be small, asymptotically. Stone [1977] notes that for probability weights with $w_i(x; X^n) \geq 0$ and $\sum_i w_i(x; X^n) = 1$, Conditions 2 and 3 are automatic, and Conditions 1, 4, and 5 are actually necessary for universal consistency, i.e., if $\widehat{\mu}$ is a universally consistent linear smoother with probability weights then Conditions 1–5 must hold.

3 Bias/variance & deriving rates

In addition to the universal consistency results derived from Stone's theorem, we also saw that when the regression function is Lipschitz (and under some additional conditions), the k-nearest neighbor estimator satisfies

$$\mathbb{E}\|\widehat{\mu} - \mu\| \lesssim \left(\frac{k}{n}\right)^{1/d} + \frac{1}{\sqrt{k}}$$

and the kernel estimator satisfies

$$\mathbb{E}\|\widehat{\mu} - \mu\| \lesssim h + \frac{1}{\sqrt{nh^d}}.$$

Here let's get a sense of how to obtain these results, focusing for simplicity on the kernel estimator.

Recall for any estimator $\widehat{\mu}$ the pointwise bias is $b(x) = \mathbb{E}\{\widehat{\mu}(x)\} - \mu(x)$ and pointwise variance is $v(x) = \mathbb{E}\Big([\widehat{\mu}(x) - \mathbb{E}\{\widehat{\mu}(x)\}]^2\Big)$, and further we have

$$\begin{aligned}\mathbb{E}\|\widehat{\mu} - \mu\| \leq \sqrt{\mathbb{E}\|\widehat{\mu} - \mu\|^2} &= \sqrt{\mathbb{E}\int \Big\{\widehat{\mu}(x) - \mu(x)\Big\}^2 \, d\mathbb{P}(x)} \\ &= \sqrt{\int \Big\{b(x)^2 + v(x)\Big\} \, d\mathbb{P}(x)}\end{aligned}$$

where the first inequality follows by Jensen's, and the last equality by the usual bias/variance decomposition of the mean squared error. Thus we will separately consider the bias and variance, and for simplicity we will focus on conditional bias/variance, given the features X^n. These arguments combined with controlling the chance of seeing empty bins (using regularity conditions in the theorem) is enough to yield the main results.

For $\widehat{\mu}$ the kernel estimator with boxcar kernel, letting $\mathcal{E}_h(x)$ denote the bad 'empty bin" event that $\sum_i \mathbb{1}(\|X_i - x\| \leq h) = 0$ and $\overline{\mathcal{E}}_h(x)$ its complement, the conditional bias is

$$\begin{aligned}\mathbb{E}\{\widehat{\mu}(x) \mid X^n\} - \mu(x) &= \mathbb{E}\left[\frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h) Y_i}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)} \, \mathbb{1}(\overline{\mathcal{E}}_h(x)) \;\Big|\; X^n\right] - \mu(x) \\ &= \mathbb{E}\left[\frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h)\{Y_i - \mu(x)\}}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)} \, \mathbb{1}(\overline{\mathcal{E}}_h(x)) \;\Big|\; X^n\right] - \mu(x)\mathbb{1}(\mathcal{E}_h(x))\end{aligned}$$

The first term is

$$\begin{aligned}\frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h)\{\mu(X_i) - \mu(x)\}}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)} \, \mathbb{1}(\overline{\mathcal{E}}_h(x)) &\leq \frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h)|\mu(X_i) - \mu(x)|}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)} \\ &\leq \frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h) L\|X_i - x\|}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)} \leq Lh\end{aligned}$$

and is the main part of the bias. The second term can be controlled by reasoning about the chance of the bad event $\mathcal{E}_h(x)$ based on the feature distribution.

The conditional variance is

$$\begin{aligned}\mathrm{var}\{\widehat{\mu}(x) \mid X^n\} &= \mathrm{var}\left[\frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h) Y_i}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)} \, \mathbb{1}(\overline{\mathcal{E}}_h(x)) \;\Big|\; X^n\right] \\ &= \frac{\sum_i \mathbb{1}(\|X_i - x\| \leq h)\mathrm{var}(Y \mid X_i)}{\{\sum_i \mathbb{1}(\|X_i - x\| \leq h)\}^2} \, \mathbb{1}(\overline{\mathcal{E}}_h(x)) \\ &\leq \frac{\sigma^2 \, \mathbb{1}(\overline{\mathcal{E}}_h(x))}{\sum_i \mathbb{1}(\|X_i - x\| \leq h)}\end{aligned}$$

If the X_i were equally spaced, we would have $\sum_i \mathbb{1}(\|X_i - x\| \leq h) = nh^d$. In a random design, $\sum_i \mathbb{1}(\|X_i - x\| \leq h) \sim \mathrm{Bin}(n, p_h(x))$ for $p_h(x) = \mathbb{P}(\|X_i - x\| \leq h)$, and one can use Lemma A.2(ii) of Devroye et al. to upper bound the expected Binomial reciprocal, along with the fact that $p_h(x) \asymp h^d$ when the distribution of X has a bounded density.

4 Tuning & optimality

Above we saw that simple estimators like k-nearest neighbors and kernel estimators are universally consistent under mild conditions (i.e., their L_2 risk converges to zero as $n \to \infty$ for any distribution whose outcomes have bounded second moments). We also saw that when the regression function is Lipschitz (and under some additional conditions), the k-nearest neighbor estimator satisfies

$$\mathbb{E}\|\widehat{\mu} - \mu\| \lesssim \left(\frac{k}{n}\right)^{1/d} + \frac{1}{\sqrt{k}}$$

and the kernel estimator satisfies

$$\mathbb{E}\|\widehat{\mu} - \mu\| \lesssim h + \frac{1}{\sqrt{nh^d}}.$$

What happens when we pick the tuning parameters k and h to minimize these errors? What is the corresponding rate of convergence?

For the k-nearest neighbor estimator we need to balance the squared bias and variance by taking

$$\left(\frac{k}{n}\right)^{2/d} \sim \frac{1}{k} \iff k^{(2+d)/d} \sim n^{2/d} \iff k \sim n^{2/(2+d)}$$

which gives

$$\mathbb{E}\|\widehat{\mu} - \mu\| \lesssim \left(\frac{k}{n}\right)^{1/d} + \frac{1}{\sqrt{k}} \lesssim n^{-1/(2+d)}.$$

Similarly, for the kernel estimator we balance squared bias and variance by taking

$$h^2 \sim \frac{1}{nh^d} \iff h^{2+d} \sim n^{-1} \iff h \sim n^{-1/(2+d)}$$

which gives

$$\mathbb{E}\|\widehat{\mu} - \mu\| \lesssim h + \frac{1}{\sqrt{nh^d}} \lesssim n^{-1/(2+d)}.$$

Are these rates optimal? Recall our minimax lower bound from last time, which told us that for the model $\mathcal{P}$ containing all distributions for which the covariate density is bounded and the regression function is Lipschitz (i.e., $s = 1$ in previous notation), we have

$$\inf_{\widehat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\|\widehat{\mu} - \mu_P\| \gtrsim n^{-1/(2+d)}.$$

Therefore the k-nearest neighbor and kernel estimators are optimal when the regression function is Lipschitz smooth.

This raises several interesting issues. First, note that the above rate, although optimal for Lipschitz smooth functions, can be quite slow depending on the dimension. For example, when $d = 20$, the optimal rate is $n^{-1/22}$, and for an estimator to achieve an error $\mathbb{E}\|\widehat{\mu} - \mu\| \leq 0.1$ we would need $n \approx 10^{22}$ samples. This is about as many stars there are estimated to be in the entire observable universe, more than all the grains of sand on Earth. What kinds of structure could help ameliorate this problem?

References

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

C. J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.