STATISTICS AND DATA SCIENCE 365 / 665
INTERMEDIATE MACHINE LEARNING

A note on the bias-variance tradeoff

In this note we give a derivation of the bias-variance tradeoff and curse of dimensionality for kernel density estimation. The analysis for kernel smoothing is similar.

The kernel density estimator is

$$\widehat{f}(x) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{1}$$

for a kernel $K(u) \geq 0$ satisfying $\int K(u)\, du = 1$ and $\int uK(u)\, du = 0$. In p-dimensions we can take the product kernel $K(u) \equiv \prod_{j=1}^{p} K(u_j)$ where $K(\cdot)$ is a 1-dimensional kernel.

To estimate the bias we calculate $\mathbb{E}\widehat{f}(x)$ as

$$\begin{aligned}
\mathbb{E}\widehat{f}(x) &= \frac{1}{nh^p} \sum_{i=1}^{n} \mathbb{E}K\left(\frac{X_i - x}{h}\right) &(2)\\
&= \frac{1}{h^p} \int K\left(\frac{u - x}{h}\right) f(u)\, du &(3)
\end{aligned}$$

since $X_i \sim F$ are iid with density f. Now make the change of variables $v = \frac{u-x}{h}$ so that $u = x + hv$ and $du = h^p dv$. Letting $h \to 0$ and assuming f is bounded and twice differentiable with bounded derivatives this gives

$$\begin{aligned}
\mathbb{E}\widehat{f}(x) &= \int K(v)\, f(x + hv)\, dv &(4)\\
&= \int K(v) \left(f(x) + hv^T \nabla f(x) + \frac{1}{2} h^2 v^T \nabla^2 f(x) v + o(h^2)\right) dv &(5)\\
&= f(x) + C(x)h^2 + o(h^2) &(6)
\end{aligned}$$

where we use the assumptions $\int K(u)\, du = 1$ and $\int uK(u)\, du = 0$ on the kernel, and that

$$C(x) = \frac{1}{2} \int K(u) u^T \nabla^2 f(x) u\, du \leq C_1 h^2 \tag{7}$$

for some constant C_1. This shows that the squared bias is of order h^4:

$$\left(\mathbb{E}\widehat{f}(x) - f(x)\right)^2 \leq C_1^2 h^4 + o(h^4). \tag{8}$$

Now we bound the variance. We use $\mathrm{Var}(x) \leq \mathbb{E}\widehat{f}(x)^2$ and calculate

$$\mathbb{E}\widehat{f}(x) \leq C_2 \frac{1}{n^2 h^{2p}} \sum_{i=1}^{n} \mathbb{E} K\left(\frac{X_i - x}{h}\right)^2 \tag{9}$$

$$= C_2 \frac{1}{nh^{2p}} \int K\left(\frac{u-x}{h}\right)^2 f(u)\, du \tag{10}$$

$$= C_2 \frac{f(x)}{nh^p} \int K(v)^2\, dv + o\left(\frac{1}{nh^p}\right) \tag{11}$$

$$= C_2 \frac{f(x)}{nh^p} + o\left(\frac{1}{nh^p}\right) \tag{12}$$

using another Taylor approximation, where we assume that $nh^p \to \infty$, with the constant C_2 changing from line to line.

To summarize, these calculations tell us that

$$\mathrm{bias}^2(x) \approx h^4 \tag{13}$$

$$\mathrm{var}(x) \approx \frac{1}{nh^p}. \tag{14}$$

Choosing h so that the squared bias and the variance are of the same order gives that

$$h \approx n^{\frac{-1}{4+p}} \tag{15}$$

$$\mathbb{E}(\widehat{f}(x) - f(x))^2 = O\left(n^{\frac{-4}{4+p}}\right). \tag{16}$$

Lower bounds tell us that this is the fastest rate the risk can decrease, under the given assumptions on the density f. Flipping this risk around, we see that the sample size n must increase exponentially in p in order to drive the risk down to a fixed level ϵ. This is the "curse of dimensionality."

Finally, note that we can think of the kernel smoothing estimator as what we get by plugging in the kernel density estimator for x and y. That is, using the "plug-in" estimate of $\widehat{f}(y \,|\, x)$ we have that

$$\widehat{m}(x) = \int y\, \widehat{f}(y \,|\, x)\, dy = \int y\, \frac{\widehat{f}(x,y)}{\widehat{f}(x)}\, dy \tag{17}$$

$$= \frac{\displaystyle\int y \frac{1}{nh^{p+1}} \sum_{i=1}^{n} K\left(\frac{Y_i - y}{h}\right) K\left(\frac{X_i - x}{h}\right) dy}{\displaystyle\frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \tag{18}$$

$$= \frac{\displaystyle\sum_{i=1}^{n} \left\{\int y \frac{1}{h} K\left(\frac{Y_i - y}{h}\right) dy\right\} K\left(\frac{X_i - x}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} = \frac{\displaystyle\sum_{i=1}^{n} Y_i\, K\left(\frac{X_i - x}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}. \tag{19}$$

This is the usual kernel smoothing estimator. This suggests that the same bias-variance decomposition holds for kernel smoothing, which is indeed the case.