11 Nearest Neighbor Methods

11.1 kth Nearest Neighbor

An alternative nonparametric method is called k-nearest neighbors or k-nn. It is simiar to kernel methods with a random and variable bandwidth. The idea is to base estimation on a fixed number of observations k which are closest to the desired point.

Suppose $X \in \mathbb{R}^q$ and we have a sample $\{X_1, ..., X_n\}$.

For any fixed point $x \in \mathbb{R}^q$, we can calculate how close each observation X_i is to x using the Euclidean distance $\|x\| = (x'x)^{1/2}$. This distance is

$$D_i = \|x - X_i\| = \left((x - X_i)'(x - X_i)\right)^{1/2}$$

This is just a simple calculation on the data set.

The order statistics for the distances D_i are $0 \leq D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(n)}$.

The observations corresponding to these order statistics are the "nearest neighbors" of x. The first nearest neighbor is the observation closest to x, the second nearest neighbor is the observation second closest, etc.

This ranks the data by how close they are to x. Imagine drawing a small ball about x and slowly inflating it. As the ball hits the first observation X_i, this is the "first nearest neighbor" of x. As the ball further inflates and hits a second observation, this observation is the second nearest neighbor.

The observations ranked by the distances, or "nearest neighbors", are $\{X_{(1)}, X_{(2)}, X_{(3)}, ..., X_{(n)}\}$.

The k'th nearest neighbor of x is $X_{(k)}$.

For a given k, let

$$R_x = \left\|X_{(k)} - x\right\| = D_{(k)}$$

denote the Euclidean distance between x and $X_{(k)}$. R_x is just the k'th order statistic on the distances D_i.

Side Comment: When X is multivariate the nearest neighbor ordering is not invariant to data scaling. Before applying nearest neighbor methods, is therefore essential that the elements of X be scaled so that they are similar and comparable across elements.

11.2 k-nn Density Estimate

Suppose $X \in \mathbb{R}^q$ has multivariate density $f(x)$ and we are estimating $f(x)$ at x.

A multivariate uniform kernel is

$$w(\|u\|) = c_q^{-1} 1(\|u\| \leq 1)$$

where

$$c_q = \frac{\pi^{q/2}}{\Gamma\left(\frac{q+2}{2}\right)}$$

is the volume of unit ball in $\mathbb{R}^q$. If $q = 1$ then $c_1 = 2$.

Treating R_x as a bandwidth and using this uniform kernel

$$\begin{aligned}\tilde{f}(x) &= \frac{1}{nR_x^q}\sum_{i=1}^{n} c_q^{-1} 1\left(\|x - X_i\| \leq R_x\right) \\ &= \frac{1}{nR_x^q}\sum_{i=1}^{n} c_q^{-1} 1\left(D_i \leq R_x\right)\end{aligned}$$

But as $R_x = D_{(k)}$ is the k'th order statistic for D_i, there are precisely k observations where $\|x - X_i\| \leq R_x$. Thus the above equals

$$\tilde{f}(x) = \frac{k}{nR_x^q c_q}$$

To compute $\tilde{f}(x)$, all you need to know is R_x.

The estimator is inversely proportional to R_x. Intuitively, if R_x is small this means that there are many observations near x, so $f(x)$ must be large, while if R_x is large this means that there are not many observations near x, so $f(x)$ must be small.

A motivation for this estimator is that the effective number of observations to estimate $\tilde{f}(x)$ is k, which is constant regardless of x. This is in contrast to the conventional kernel estimator, where the effective number of observations varies with x.

While the traditional k-nn estimator used a uniform kernel, smooth kernels can also be used. A smooth k-nn estimator is

$$\tilde{f}(x) = \frac{1}{nR_x^q}\sum_{i=1}^{n} w\left(\frac{\|x - X_i\|}{R_x}\right)$$

where w is a kernel weight function such that

$$\int_{\mathbb{R}^q} w\left(\|u\|\right)(du) = 1.$$

In this case the estimator does not simplify to a function of R_x only

The analysis of k-nn estimates are complicated by the fact that R_x is random.

The solution is to calculate the bias and variance of $\hat{f}(x)$ conditional on R_x, which is similar to treating R_x as fixed. It turns out that the conditional bias and variance are identical to those of the standard kernel estimator:

$$\begin{aligned}E\left(\tilde{f}(x) \mid R_q\right) &\simeq f(x) + \frac{\kappa_2(w)\nabla^2 f(x) R_x^2}{2} \\ var\left(\tilde{f}(x) \mid R_q\right) &\simeq \frac{R\left(w\right) f(x)}{nR_x^q}.\end{aligned}$$

We can then approximate the unconditional bias and variance by taking expectations:

$$\begin{aligned} E\left(\tilde{f}(x)\right) &\simeq f(x) + \frac{\kappa_2(w)\nabla^2 f(x)}{2} E\left(R_x^2\right) \\ var\left(\tilde{f}(x)\right) &\simeq \frac{R(w) f(x)}{n} E\left(R_x^{-q}\right) \end{aligned}$$

We see that to evaluate these expressions we need the moments of $R_x = D_{(k)}$ the k'th order statistic for D_i. The distribution function for order statistics is well known. Asymptotic moments for the order statistics were found by Mack and Rosenblatt (Journal of Multivariate Analysis, 1979):

$$E\left(R_x^{\lambda}\right) \simeq \left(\frac{k/n}{c_q f(x)}\right)^{\lambda/q}$$

This depends on the ratio k/n and the density $f(x)$ at x. Thus

$$\begin{aligned} E\left(R_x^2\right) &\simeq \left(\frac{k}{n c_q f(x)}\right)^{2/q} \\ E\left(R_x^{-q}\right) &\simeq \frac{c_q f(x) n}{k} \end{aligned}$$

Substituting,

$$\begin{aligned} Bias\left(\tilde{f}(x)\right) &\simeq \frac{\kappa_2(w)\nabla^2 f(x)}{2} \left(\frac{k}{n c_q f(x)}\right)^{2/q} \\ &= \frac{\kappa_2(w)\nabla^2 f(x)}{2 \left(c_q f(x)\right)^{2/q}} \left(\frac{k}{n}\right)^{2/q} \end{aligned}$$

$$\begin{aligned} var\left(\tilde{f}(x)\right) &\simeq \frac{R(w) f(x)}{n} \frac{c_q f(x)}{k} n \\ &= \frac{R(w) c_q f(x)^2}{k} \end{aligned}$$

For k-nn estimation, the integer k is similar to the bandwidth h for kernel density estimation, except that we need $k \to \infty$ and $k/n \to 0$ as $n \to \infty$.

The MSE is of order

$$MSE\left(\tilde{f}(x)\right) = O\left(\left(\frac{k}{n}\right)^{4/q} + \frac{1}{k}\right)$$

This is minimized by setting

$$k \sim n^{4/(4+q)}.$$

The optimal rate for the MSE is

$$MSE\left(\tilde{f}(x)\right) = O\left(n^{-4/(4+q)}\right)$$

which is the same as for kernel density estimation with a second-order kernel.

Kernel estimates $\hat{f}$ and k-nn estimates $\tilde{f}$ behave differently in the tails of $f(x)$ (where $f(x)$ is small). The contrast is

$$\begin{aligned} Bias\left(\hat{f}(x)\right) &\simeq \nabla^2 f(x) \\ Bias\left(\tilde{f}(x)\right) &\simeq \frac{\nabla^2 f(x)}{f(x)^{2/q}} \end{aligned}$$

$$\begin{aligned} var\left(\hat{f}(x)\right) &\simeq f(x) \\ var\left(\tilde{f}(x)\right) &\simeq f(x)^2 \end{aligned}$$

In the tails, where $f(x)$ is small, $\tilde{f}(x)$ will have larger bias but smaller variance than $\hat{f}(x)$. This is because the k-nn estimate uses more effective observations than the kernel estimator. It is difficult to rank one estimator versus the other based on this comparison. Another way of viewing this is that in the tails $\tilde{f}(x)$ will tend to be smoother than $\hat{f}(x)$.

11.3 Regression

Nearest neighbor methods are more typically used for regression than for density estimation.

The regression model is

$$\begin{aligned} y_i &= g(X_i) + e_i \\ E(e_i \mid X_i) &= 0 \end{aligned}$$

The classic k-nn estimate of $g(x)$ is

$$\tilde{g}(x) = \frac{1}{k}\sum_{i=1}^{n} 1(\|x - X_i\| \leq R_x)\, y_i$$

This is the average value of y_i among the observations which are the k nearest neighbors of x.

A smooth k-nn estimator is

$$\tilde{g}(x) = \frac{\sum_{i=1}^{n} w\left(\frac{\|x - X_i\|}{R_x}\right) y_i}{\sum_{i=1}^{n} w\left(\frac{\|x - X_i\|}{R_x}\right)},$$

a weighted average of the k nearest neighbors.

The asymptotic analysis is the same as for density estimation. Conditional on R_x, the bias and variance are approximately as for NW regression. The conditional bias is proportional to R_x^2 and

the variance to $1/nR_x^q$. Taking unconditional expecations and using the formula for the moments of R_x give expressions for the bias and variance of $\tilde{g}(x)$. The optimal rate is $k \sim n^{4/(4+q)}$ and the optimal convergence rate is the same as for NW estimation.

As for density estimation, in the tails of the density of X, the bias of the k-nn estimator is larger, and the variance smaller, than the NW estimator $\hat{g}(x)$. Since the effective number of observations k is held constant across x, $\tilde{g}(x)$ is smoother than $\hat{g}(x)$ in the tails.

11.4 Local Linear k-nn Regression

As pointed out by Li and Racine, local linear esitmation can be combined with the nearest neighbor method.

A simple estimator (corresonding to a uniform kernel) is to take the k observations "nearest" to x, and fit a linear regression of y_i on X_i using these observations.

A smooth local linear k-nn estimator fits a weighted linear regression

11.5 Cross-Validation

To use nearest neighbor methods, the integer k must be selected. This is similar to bandwidth selection, although here k is discrete, not continuous.

K.C. Li (Annals of Statistics, 1987) showed that for the $k-$nn regression estimator under conditional homoskedasticity, it is asymptotically optimal to pick k by Mallows, Generalized CV, or CV. Andrews (Journal of Econometrics, 1991) generalized this result to the case of heteroskedasticity, and showed that CV is asymptotically optimal. The CV criterion is

$$CV(k) = \sum_{i=1}^{n} (y_i - \tilde{g}_{-i}(X_i))^2$$

and $\tilde{g}_{-i}(X_i)$ is the leave-one-out k-nn estimator of $g(X_i)$. The method is to select k by minimizing $CV(k)$. As k is discrete, this amounts to computing $CV(k)$ for a set of values for k, and finding the minimizing value.