IEOR 265 – Lecture 6
Local Linear Regression

1 Local Linear Regression

Consider a regression model

$$y = f(x) + \epsilon$$

in which $f(\cdot)$ is known to be highly nonlinear but of unknown structure. A nonparametric approach is natural, and one nonparametric method is known as local linear regression (LLR). The idea of this method is that if $f(\cdot)$ has sufficient smoothness (say twice-differentiable), then the model will look linear in small regions of input-space. Suppose that we consider points in input space nearby x_0, then intuitively our model looks like

$$y = \beta_0[x_0] + \sum_{j=1}^{p} \beta_j[x_0] \cdot (x^j - x_0^j) + \epsilon$$

for x near x_0 (e.g., $\|x - x_0\| \leq h$ for some small $h > 0$). The square brackets $[x_0]$ are used to represent the fact that the value of β will vary for different values of x_0.

The idea of a neighborhood of radius h is central to LLR. It is customary in statistics to call this h the *bandwidth*. In this method, we select points within a radius of h from x_0. Furthermore, we can weight the points accordingly so that points closer to x_0 are given more weight than those points further from x_0. To do this, we define a kernel function $K(u) : \mathbb{R} \to \mathbb{R}$ which has the properties

1. Finite Support – $K(u) = 0$ for $|u| \geq 1$;
2. Even Symmetry – $K(u) = K(-u)$;
3. Positive Values – $K(u) > 0$ for $|u| < 1$.

A canonical example is the Epanechnikov kernel

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2), & \text{for } |u| < 1 \\ 0, & \text{otherwise} \end{cases}$$

It turns out that the particular shape of the kernel function is not as important as the bandwidth h. If we choose a large h, then the local linear assumption is not accurate. On the other hand, if we choose a very small h, then the estimate will not be accurate because only a few data points will be considered. It turns out that this tradeoff in the value of h is a manifestation of the bias-variance tradeoff.

Before we discuss this tradeoff in more detail, we describe the LLR. The idea is to perform a weighted-variant of OLS by using a kernel function and a bandwidth h to provide the weighting. The LLR estimate $\hat{\beta}_0[x_0], \hat{\beta}[x_0]$ is given by the minimizer to the following optimization

$$\begin{bmatrix} \hat{\beta}_0[x_0] \\ \hat{\beta}[x_0] \end{bmatrix} = \arg\min_{\beta_0,\beta} \sum_{i=1}^{n} K(\|x_i - x_0\|/h) \cdot (y_i - \beta_0 - (x_i - x_0)'\beta)^2.$$

Now if we define a weighting matrix

$$W_h = \mathsf{diag}\left(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\right),$$

then we can rewrite this optimization as

$$\begin{bmatrix} \hat{\beta}_0[x_0] \\ \hat{\beta}[x_0] \end{bmatrix} = \arg\min_{\beta_0,\beta} \left\| W_h^{1/2} \left(Y - \begin{bmatrix} \mathbf{1}_n & X_0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \right) \right\|_2^2,$$

where $\mathbf{1}_n$ is a real-valued vector of all ones and of length dimension n and $X_0 = X - x_0'\mathbf{1}_n$. This is identical to the OLS optimization, and so we can use that answer to conclude that

$$\begin{bmatrix} \hat{\beta}_0[x_0] \\ \hat{\beta}[x_0] \end{bmatrix} = (\begin{bmatrix} \mathbf{1}_n & X_0 \end{bmatrix}' W_h \begin{bmatrix} \mathbf{1}_n & X_0 \end{bmatrix})^{-1} (\begin{bmatrix} \mathbf{1}_n & X_0 \end{bmatrix}' W_h Y).$$

2 Bias and Variance of LLR

Next, we describe the bias and variance of LLR. The heuristic derivations are more complicated, and so we do not include them. It can be shown that the bias and variance of an LLR estimate is

$$\begin{aligned} \mathsf{bias}(\hat{\beta}_0[x_0]) &= O(h^2) \\ \mathsf{var}(\hat{\beta}_0[x_0]) &= O_p\left(\frac{1}{nh^d}\right), \end{aligned}$$

where h is the bandwidth and d is the dimensionality of x_i (i.e., $x_i \in \mathbb{R}^d$). Note that we have described the bias and variance for the estimate of β_0 and not that of β. The bias and variance for the estimate of β are slightly different. The interesting aspect is that these expressions show how the choice of the bandwidth h affects bias and variance. A large bandwidth means that we have high bias and low variance, whereas a small bandwidth means that we have a low bias but high variance.

When implementing LLR, we need to pick the bandwidth h in some optimal sense that trades off

these competing effects. We can, at least analytically, pick the asymptotic rate of h:

$$\begin{aligned} h &= \arg\min_h \left(c_1 h^4 + c_2 \frac{1}{nh^d} \right) \\ \Rightarrow \tilde{c}_1 h^3 + \tilde{c}_2 \frac{1}{nh^{d+1}} &= 0 \\ \Rightarrow \bar{c}_1 h^{d+4} &= \bar{c}_2 \frac{1}{n} \\ \Rightarrow h &= O(n^{-1/(d+4)}). \end{aligned}$$

The resulting optimal rate is that

$$\beta_0 = \beta + O_p(n^{-2/(d+4)}).$$

Note that this is slower than the rate of OLS (i.e, $O_p(1/\sqrt{n})$) for any $d \geq 1$, and is an example of the statistical penalty that occurs when using nonparametric methods. The situation is in fact even worse: There is a "curse of dimensionality" that occurs, because the convergence rate gets exponentially worse as the ambient dimension d increases. It turns out that if the model has additional structure, then one can improve upon this "curse of dimensionality" and get better rates of convergence when using LLR.

3 Manifold Structure

Imagine a generalization of the collinearity model previously considered. Specifically, assume that the x_i lie on an embedded submanifold $\mathcal{M}$ with dimension $d < p$, where this manifold is unknown *a priori*. Suppose that the system has model $y_i = f(x_i) + \epsilon_i$, where $f(\cdot)$ is an unknown nonlinear function. The reason that this situation is interesting is that the LLR estimate converged at rate $O_p(n^{-2/(p+4)})$, but if we knew the manifold then we could do a coordinate change into a lower-dimensional space and then the LLR estimate would converge at rate $O_p(n^{-2/(d+4)})$.

Even though this manifold is unknown, we could imagine that if we were able to somehow learn this manifold and then incorporate this knowledge into our estimator, then we could achieve the faster convergence rate. This is in fact the idea behind the nonparametric exterior derivative estimator (NEDE), which is defined as

$$\begin{bmatrix} \hat{\beta}_0[x_0] \\ \hat{\beta}[x_0] \end{bmatrix} = \arg\min_{\beta_0, \beta} \left\| W_h^{1/2} \left(Y - \begin{bmatrix} 1_n & X_0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \right) \right\|_2^2 + \lambda \|\Pi \beta\|_2^2,$$

where $X_0 = X - x_0' 1_n$, Π is a projection matrix that projects onto the $(p-d)$ smallest eigenvectors of the sample local covariance matrix $\frac{1}{nh^{d+2}} X_0' W_h X_0$, and

$$W_h = \text{diag}\left(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h) \right).$$

It can be shown that the error in this estimate converges at rate $O_p(n^{-2/(d+4)})$, even though the regression is being computed for coefficients that lie within a p-dimensional space. Furthermore,

it can be shown that $\hat{\beta}[x_0]$ is a consistent estimate of the exterior derivative of f at x_0 (i.e., $df|_{x_0}$). This improved convergence rate can be very useful if $d \ll p$.

The idea of lower-dimensional structure either in a hyperplane or manifold context is an important abstract structure. It is important because there are many methods that can exploit such structure to provide improved estimation.

3.1 Collinearity and Sparsity

In some models, one might have both collinearity and sparsity. One approach to this situation is the *elastic net*, which is

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 + \mu\|\beta\|_1.$$

An alternative approach might be the Lasso Exterior Derivative Estimator (LEDE) estimator

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\Pi\beta\|_2^2 + \mu\|\beta\|_1,$$

where Π is a projection matrix that projects onto the $(p-d)$ smallest eigenvectors of the sample covariance matrix $\frac{1}{n}X'X$.

A further generalization of this idea is when there is manifold structure and sparsity: The Non-parametric Lasso Exterior Derivative Estimator (NLEDE) estimator is

$$\begin{bmatrix}\hat{\beta}_0[x_0] \\ \hat{\beta}[x_0]\end{bmatrix} = \arg\min_{\beta_0,\beta} \left\| W_h^{1/2}\left(Y - \begin{bmatrix}\mathbf{1}_n & X_0\end{bmatrix}\begin{bmatrix}\beta_0 \\ \beta\end{bmatrix}\right)\right\|_2^2 + \lambda\|\Pi\beta\|_2^2 + \mu\|\beta\|_1,$$

where $X_0 = X - x_0'1_n$, Π is a projection matrix that projects onto the $(p-d)$ smallest eigenvectors of the sample local covariance matrix $\frac{1}{nh^{d+2}}X_0'W_hX_0$, and

$$W_h = \mathsf{diag}\left(K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\right).$$