36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 18: October 10

Lecturer: Siva Balakrishnan

18.1 Review and Outline

In the last class we discussed linear regression:

1. Multiple Regression
2. Statistical Modeling
3. Non-parametric Regression

Today we will continue our discussion of non-parametric regression. Particularly, we focus on understanding the bias-variance tradeoff in non-parametric regression.

18.2 Non-parametric Regression

Recall that broadly in regression our goal is to estimate the regression function $r(x) = \mathbb{E}[Y|X = x]$. Unlike the CDF which we could estimate with no assumptions about the distribution, here we will need *smoothness* assumptions, i.e. we will need to assume that $r(x)$ is a smooth function of x. This allows us to gain statistical strength by averaging near by points.

Suppose we construct an estimate $\widehat{r}(x)$. Then a natural measure of how well we do is the squared loss, except since these are functions this is called the *integrated* squared loss, i.e.:

$$L(\widehat{r}, r) = \int (\widehat{r}(x) - r(x))^2 dx.$$

The risk is then just the expected loss, i.e.:

$$R(\widehat{r}, r) = \mathbb{E}\left(\int (\widehat{r}(x) - r(x))^2 dx\right).$$

As in the case of point estimation we have a bias variance decomposition. First we define the point-wise bias:

$$b(x) = \mathbb{E}(\widehat{r}(x)) - r(x),$$

and the point-wise variance:

$$v(x) = \mathbb{E}(\widehat{r}(x) - \mathbb{E}(\widehat{r}(x)))^2.$$

Now, as before we can verify that:

$$R(\widehat{r}, r) = \int b^2(x)dx + \int v(x)dx.$$

A natural strategy in non-parametric regression is to locally average the data, i.e. our estimate of the regression function at any point will be the average of the Y values in a small neighborhood of the point.

The width of this neighborhood will determine the bias and variance. Too large a neighborhood will result in high bias and low variance (this is called oversmoothing) and too small a neighborhood will result in low bias but large variance (this is known as undersmoothing).

18.3 Kernel Regression

One of the most basic ways of doing non-parametric regression is called kernel regression. We will analyze kernel regression when we only have one covariate. The general case is not very different.

Here the estimator is defined as:

$$\widehat{r}(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where the weights assign more importance to points near x. This is called a kernel regressor when the weights are chosen according to a kernel, i.e. we have weights:

$$w_i(x) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)},$$

where h controls the amount of smoothing. It is called the bandwidth. As a typical common example we have the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

18.4 Analysis of Kernel Regression

In this section we will provide a simple analysis of kernel regression. We will do so under various (strong) assumptions. We make these assumptions so that we can prove our main result in this lecture.

We assume that

$$y_i = r(x_i) + \epsilon_i,$$

where:

1. **Assumptions about the design:** We will assume that x_i is one-dimensional, and equally spaced on $[0, 1]$. In general, we do not need that the design is equally spaced but intutively we do need to ensure that we see some points in the vicinity of each point where f is non-zero.
2. **Assumptions about the regression function:** We will assume that the function $r(x) = \mathbb{E}[Y|X = x]$ is L-Lipschitz, i.e. that there is some constant L such that:
$$\left|\frac{d}{dx}r(x)\right| \leq L.$$
3. **Assumptions about the noise:** We will assume that the noise is i.i.d and that $\mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2$.
4. **Assumptions about the kernel:** We will assume that the kernel is the *spherical kernel*:
$$K(x) = \mathbb{I}(-1 \leq x \leq 1).$$
Again the result we prove holds for a much broader class of kernels but we make this assumption to simplify the proof.

The main result: Suppose that the bandwidth $h \geq 1/n$, then the point-wise bias of the kernel regression estimator:

$$b(x) = \mathbb{E}[\widehat{r}(x)] - r(x) \leq Lh,$$

and the point-wise variance:

$$\mathrm{Var}(\widehat{r}(x)) = \mathbb{E}(\widehat{r}(x) - \mathbb{E}(\widehat{r}(x)))^2 \leq \frac{\sigma^2}{nh}.$$

Before, we prove the theorem let us observe that we can calculate the integrated risk:

$$R(\widehat{r}, r) = \int b^2(x)dx + \int v(x)dx \leq L^2h^2 + \frac{\sigma^2}{nh},$$

a natural strategy would be to choose the bandwidth to minimize this expression. Choosing

$$h = \left(\frac{\sigma^2}{2nL^2}\right)^{1/3},$$

we obtain that

$$R(\widehat{r}, r) \leq \frac{2L^{2/3}\sigma^{4/3}}{n^{2/3}} = 2\left(\frac{L\sigma^2}{n}\right)^{2/3}.$$

This result already reveals something fundamental about non-parametrics. We see that with an optimally chosen bandwidth, the MISE decreases to 0 at rate $n^{-2/3}$. By comparison, most parametric estimators converge at rate n^{-1}. The slower rate of convergence is the price we pay for being nonparametric. The formula for the optimal bandwidth is not useful in practice since it depends on the unknown Lipschitz constant L, and the typically unknown noise variance.

With all of these preliminaries in place let us prove the result:

Bounding the bias: The bias is given by:

$$\begin{aligned} b(x) = |\mathbb{E}\widehat{r}(x) - r(x)| &= \left|\mathbb{E}\left[\sum_{i=1}^{n}(w_i(X_i)(Y_i - r(x)))\right]\right| \\ &= \left|\sum_{i=1}^{n}(w_i(X_i)(r(X_i) - r(x)))\right| \leq \sum_{i=1}^{n} w(X_i)|r(X_i) - r(x)| \\ &\leq Lh\sum_{i=1}^{n} w(X_i) = Lh. \end{aligned}$$

At a high-level, the kernel regressor aggregates Y values from nearby points, and the bias just captures how much the true function can change over this region.

Bounding the variance: We note first that each weight is upper bounded as:

$$w_i(X_i) \leq \frac{1}{nh}.$$

Returning to the variance we obtain:

$$\begin{aligned} v(x) = \mathbb{E}(\widehat{r}(x) - \mathbb{E}(\widehat{r}(x)))^2 &= \mathbb{E}\left(\sum_{i=1}^{n}(w_i(X_i)(Y_i - f(X_i)))\right)^2 \\ &= \mathbb{E}\left(\sum_{i=1}^{n}\epsilon_i w_i(X_i)\right)^2 \\ &= \sum_{i=1}^{n} w_i(X_i)^2\mathbb{E}(\epsilon_i^2) = \sigma^2\sum_{i=1}^{n} w_i(X_i)^2 \\ &\leq \sigma^2 \max_{i=1}^{n} w_i(X_i)\sum_{i=1}^{n} w_i(X_i) \leq \frac{\sigma^2}{nh}. \end{aligned}$$

This completes our analysis.

18.5 The general case

More generally, suppose that the β^{th} derivative of $r(x)$ is bounded, and we are in d-dimensions. In this case the bias will be roughly:

$$b^2(x) \approx h^{2\beta},$$

and the variance:

$$v(x) \approx \frac{1}{nh^d},$$

and balancing these will lead to the rate of convergence:

$$R(\widehat{r}, r) \approx n^{-2\beta/(2\beta+d)}.$$

This reveals another crucial feature of non-parametrics. In linear regression, the rate of convergence is typically something like:

$$R(\widehat{\beta}, \beta) \approx \frac{d}{n}.$$

In both cases, the situation gets worse as d increases, however in non-parametrics the situation gets *exponentially* worse. This is often colloquially referred to as the *curse of dimensionality.*