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

Lecture 19: October 12

Lecturer: Siva Balakrishnan

19.1 Review and Outline

In the last class we discussed linear regression:

1. Non-parametric regression for Lipschitz functions in 1D
2. General non-parametric regression

Today we will discuss non-parametric density estimation. This material is also in Chapter 20 of the Wasserman book.

19.2 Density Estimation

The task in density estimation is that we receive i.i.d. samples:

$$X_1, \ldots, X_n \sim f$$

where f is some (smooth) density, and would like to estimate f. Our earlier lectures on point estimation essentially addressed the parametric analogue of this problem. Concretely, we have seen how to estimate θ under the hypothesis that:

$$X_1, \ldots, X_n \sim f_\theta,$$

and today we will study the non-parametric analogue of this problem. We will focus mostly on the one-dimensional case.

19.3 Histograms

Suppose that $X_1, \ldots, X_n \sim f$, where f is a density supported on $[0, 1]$ (this restriction is not critical). A natural estimator in this context is to bin the samples in a certain way, i.e., we define the bins $B_1, \ldots, B_m$ to be:

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ldots, B_m = \left[\frac{m-1}{m}, 1\right].$$

Let $\nu_1, \ldots, \nu_m$ be the number of samples in each bin. We can then estimate the *probability mass* of each bin as:

$$\widehat{p}_1 = \frac{\nu_1}{n},$$
$$\vdots$$
$$\widehat{p}_m = \frac{\nu_m}{n}.$$

To estimate the density (our original goal) we use for an x that lies in the i^{th} bin:

$$\widehat{f}(x) = \frac{m\nu_i}{n}.$$

We can also write this in terms of the *bin-width* $h = \frac{1}{m}$ as:

$$\widehat{f}(x) = \frac{\widehat{p}_i}{h}.$$

We can also express this as:

$$\widehat{f}(x) = \sum_{i=1}^{m} \frac{\widehat{p}_i}{h} \mathbb{I}(x \in B_i).$$

As a sort of basic sanity check we can see that if $x \in B_i$ then

$$\mathbb{E}[\widehat{f}(x)] = \frac{\mathbb{E}[\widehat{p}_i]}{h} = \frac{p_i}{h} = \frac{\int_{x \in B_i} f(x)dx}{h} \approx \frac{hf(x)}{h} = f(x).$$

The $\approx$ comes from assuming that the function f does not change much over the bin if h is small. You might already be able to smell a bias-variance tradeoff.

19.4 Analyzing the histogram

Once again we will use the integrated squared loss:

$$R(\widehat{f}, f) = \int_x b^2(x)dx + \int_x v(x)dx,$$

where $b(x)$ is the bias of our estimator at x and $v(x)$ is the variance of our estimator at x. We will assume that the unknown density is L-Lipschitz. Observe that unlike in the non-parametric regression case, we need very few assumptions for density estimation.

Our goal will be to show that:

$$R(\widehat{f}, f) \leq (Lh)^2 + \left[\frac{1}{nh} + \frac{L}{n}\right]$$

Before we prove this, let us develop some consequences: the optimal bandwidth is

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

and this gives us that the risk:

$$R(\widehat{f}, f) \leq 4\left(\frac{L}{n}\right)^{2/3} + \frac{L}{n} \leq 5\left(\frac{L}{n}\right)^{2/3},$$

for sufficiently large n (since L is a constant). We observe that the rate again is slower than the usual $1/n$ rate in parametric problems. Coincidentally, this is roughly the same rate as in regression where $\mathbb{E}[Y|X = x]$ is L-Lipschitz.

Bounding the Bias: Recall that:

$$b(x) = \mathbb{E}[\widehat{f}(x)] - f(x).$$

Suppose that $x \in B_i$ then

$$\begin{aligned} b(x) &= \frac{p_i}{h} - f(x) \\ &= \frac{1}{h}\left[\int_{B_i} (f(u) - f(x))du\right] \\ &\leq \frac{1}{h}\int_{B_i} Lh du = Lh. \end{aligned}$$

Bounding the Variance: We can see that:

$$\begin{aligned} v(x) &= \mathbb{E}(\widehat{f}(x) - \mathbb{E}\widehat{f}(x))^2 \\ &= \frac{\mathbb{E}(\widehat{p}_i - p_i)^2}{h^2} = \frac{p_i(1-p_i)}{nh^2} \\ &\leq \frac{p_i}{nh^2} = \frac{1}{nh^2}\int_{B_i} f(u)du \\ &\leq \frac{1}{nh^2}\int_{B_i} (f(x) + Lh)du \\ &= \frac{f(x) + Lh}{nh} \end{aligned}$$

Integrating this we obtain the integrated variance:

$$\int_x v(x)dx \leq \frac{1}{nh} + \frac{L}{n},$$

as desired.

19.5 The general case

As in the regression setting one can ask what the rate of convergence is if f is β-times differentiable, and we are estimating a d-dimensional density.

In this case, as in the regression case, the optimal rate is $n^{-2\beta/(2\beta+d)}$. Somewhat surprisingly this rate is not achieved by histograms, and we more generally need to use *kernel density estimators*. We will discuss these in the next lecture.

For now, let us again reflect on the curse of dimensionality. Suppose we fixed $\beta = 1$ (i.e. Lipschitz densities) and then said how many samples do we need to get a squared error of 0.1 in d dimensions.

We would solve the expression $n^{-2/(2+d)} \leq 0.1$, i.e., we need:

$$\log_{10} n \geq \frac{2+d}{2},$$

or in other words:

$$n \geq 10^{1+d/2}.$$

This is astronomical for large d, i.e., this gives:

$$\begin{aligned} n &\geq 32 \text{ if } d = 1 \\ n &\geq 100 \text{ if } d = 2 \\ &\vdots \\ n &\geq 10^6 \text{ if } d = 10. \end{aligned}$$

Roughly, a million points in 10-dimensions is equivalent to 32 points in 1D. In many problems we have hundreds or thousands (or many more) features, and you can see immediately that non-parametric methods will fail miserably in these settings.