Lecture # 22

- To summarize from last time, if we consider estimating the density f_X of a r.v. on $[0,1]$ with a histogram with uniform bins of width h, then for an iid sample of size n from X, we have the following error:

$$MSE = \text{Bias}^2 + \text{Variance}$$

$$= \int_0^1 \left[\mathbb{E}\left(\hat{f}_n(x)\right) - f_X(x) \right]^2 dx + \int_0^1 \mathbb{E}\left(\hat{f}_n(x) - \mathbb{E}\left(\hat{f}_n(x)\right) \right)^2 dx$$

$$\approx \frac{h^2}{12} \cdot \int_0^1 \left[f_X'(x) \right]^2 dx + \frac{1}{h} \cdot \frac{1}{n}$$

- So, small h (tiny bins) favor low bias at the expense of high variance. Vice versa for large bins.

Q: What is the optimal h? We can phrase this as a calculus problem, because $g(h) := \frac{h^2}{12} \cdot \int_0^1 \left[f_X'(x) \right]^2 dx + h^{-1} \cdot \frac{1}{n}$ is smooth in h (away from $h = 0$). Let's solve $g'(h) = 0$:

$$g'(h) = \frac{h}{6} \cdot \int_0^1 \left[f_X'(x) \right]^2 dx - h^{-2} \cdot \frac{1}{n} = 0$$

$$\Leftrightarrow \quad h^3 \cdot \int_0^1 \left[f_X'(x) \right]^2 dx - \frac{6}{n} = 0$$

$$\hookleftarrow \quad h = \left(\frac{6}{n \cdot \int_0^1 \left[f_X'(x) \right]^2 dx} \right)^{1/3}$$

$$= \frac{1}{n^{1/3}} \cdot \left(\frac{6}{\int_0^1 \left[f_X'(x) \right]^2 dx} \right)^{1/3}$$

. So, as a function of the sample size, h should scale like $h \sim n^{-\frac{1}{3}}$ to have optimal error. Plugging this back into the MISE gives an optimal MISE of

$$MISE \approx n^{-\frac{2}{3}} \cdot \left(\frac{3}{4} \right)^{2/3} \cdot \left[\int_0^1 \left[f_X'(x) \right]^2 dx \right]^{1/3}$$

. So, as $n \to \infty$, the error $\to 0$ at a rate of $n^{-\frac{2}{3}}$.

Q: Is this the best any estimator for f_X can do? No! In fact, a simple extension of histogram estimation improves on that convergence rate: kernel density estimation (KDE).

Defn: A function $K : \mathbb{R} \to \mathbb{R}$ is a kernel if:

(1.) K is infinitely differentiable.

(2) $\int_{\mathbb{R}} K(x) dx = 1$

(3.) $\int_{\mathbb{R}} x K(x) dx = 0$

(4.) $\int_{\mathbb{R}} x^2 K(x)^2 dx = \sigma^2 > 0.$

ex: $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ or more generally

$$K_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

is the Gaussian kernel.

ex: $K(x) = \begin{cases} \frac{\frac{3}{4}\left(1-\frac{x^2}{5}\right)}{\sqrt{5}}, & |x|<\sqrt{5} \\ 0, & \text{else} \end{cases}$

ex: One can take $\tilde{K}(x) = \begin{cases} 1, & |x| \le \frac{1}{2} \\ 0, & \text{else} \end{cases}$

and mollify/smooth it to be C^∞ while preserving its area under the curve and symmetry.

in certain planes

The idea of KDE is to put a kernel at each sample, instead of a rectangle as in a histogram. So, KDE can be understood as even more data adaptive than histogram estimation.

Defn: Let K be a kernel. Given a sample $\{x_i\}_{i=1}^n \subset \mathbb{R}$ the KDE associated to the sample, K, and bandwidth $h>0$ is $\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} h^{-1} \cdot K\left(\frac{x-x_i}{h}\right)$.