Lecture #21

- Linear regression has clear advantages: - simple
  - easy to understand
  - elegant interpretation that is easily generalized (e.g. ridge regression, LASSO).
- However, not all data is well-modeled by a linear fit. In this sense, linear regression is often (not under our simple linear model, of course) biased.
- Non-parametric approaches to regression try to avoid this by allowing for more complex techniques of modeling. However, one must balance fitting well (e.g. low bias) with the ability to avoid overfitting to the noise (low variance).
- Let us start/return to a classical nonparametric problem: given $\{X_i\}_{i=1}^n \overset{iid}{\sim} X$, use the data to approximate X, through its density/cdf.
- We recall that the empirical c.d.f. is

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x - x_i),$$

where $\mathbb{1}(y) = \begin{cases} 1, & y \geq 0 \\ 0, & y < 0 \end{cases}$.

- We now consider the harder problem of density estimation. One natural approach is to construct an (empirical) histogram , which may be interpreted as an estimate on the true density f_X.
- Let us proceed under the following assumptions: • X takes values on $[0,1]$.
  • X admits density f_X.

- We construct an estimate for f_X by (1.) Binning the data
  (2.) Putting a histogram over the binned data.

- More precisely, let us fix an integer N and define $B_i = \left[\frac{i-1}{N}, \frac{i}{N}\right)$ for $i = 1, 2, \ldots, N$.
  Each bin is constructed to have width $h = \frac{1}{N}$.

- Let $V_j = |\{X_i \text{ s.t. } X_i \in B_j\}|$ be the number of observations that land in B_j.

Notice that by construction, $\sum_{j=1}^{N} V_j = n$.

$$\Leftrightarrow \sum_{j=1}^{N} \frac{V_j}{n} = 1$$

- This suggests defining $\hat{\rho}_j = \frac{V_j}{n}$ and letting $\hat{f}_n(x) = \begin{cases} \hat{\rho}_1/h, & x \in B_1 \\ \hat{\rho}_2/h, & x \in B_2 \\ \vdots \\ \hat{\rho}_N/h, & x \in B_N \end{cases}$

$$= \sum_{j=1}^{N} \frac{\hat{\rho}_j}{h} \cdot \mathbb{1}_{B_j}(x)$$

Then notice this defines a density: $\int_{\mathbb{R}} \hat{f}_n(x)\,dx$

$$= \int_{\mathbb{R}} \sum_{j=1}^{N} \frac{\hat{\rho}_j}{h} \cdot \mathbb{1}_{B_j}(x)\,dx$$

$$= \sum_{j=1}^{N} \frac{\hat{\rho}_j}{h} \cdot \int_{\mathbb{R}} \mathbb{1}_{B_j}(x)\,dx$$

$$= \sum_{j=1}^{N} \frac{\hat{\rho}_j}{h} \cdot h$$

$$= \sum_{j=1}^{N} \hat{p}_j$$
$$= 1.$$

Q: In what sense is $\hat{f}_n(x)$ a good estimate for $f_X(x)$? Well,

$$\mathbb{E}\left(\hat{f}_n(x)\right) = \mathbb{E}\left(\sum_{j=1}^{N} \frac{\hat{p}_j}{h} \cdot \mathbb{1}_{B_j}(x)\right)$$

$$= \frac{1}{h} \cdot \mathbb{E}(\hat{p}_j), \quad \text{where } x \in B_j$$

$$= \frac{1}{h} \cdot \int_{B_j} f_X(y)\, dy$$

Exercise: $\mathbb{E}(\hat{p}_j) = \int_{B_j} f_X(y)\,dy$

. Now, if $h \approx 0$, then as long as f_X is smooth, $\int_{B_j} f_X(y)dy \approx f(x) \cdot h$,

since $x \in B_j$, and hence for $h \approx 0$, $\mathbb{E}(\hat{f}_n(x)) \approx f(x)$!

. The above argument suggests we should let $h \to 0^+$, so that ↖ holds. But, there is an unfortunate downside to this: our histogram becomes a bunch of "spikes" as $h \to 0^+$!

. The underlying phenomenon in how to think about h is a bias-variance tradeoff.

Proposition: Fix X and N, with $f(x)$, $\hat{f}_n(x)$, $\{x_i\}_{i=1}^{n}$ as above. Then:

(i) $\mathbb{E}(\hat{f}_n(x)) = \frac{1}{h} \cdot \int_{B_j} f_X(y)dy$, $x \in B_j$

(ii) $\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh}\left(\int_{B_j} f_X(y)dy\right)\left(1 - \int_{B_j} f_X(y)dy\right)$

Proof: We already did (i). Proving (ii) is similar if it is left as an exercise.

- To simplify notation, let $\rho_j = \int_{B_j} f_X(x)dx$, so that $\hat{\rho}_j$ is a (unbiased) estimator for ρ_j.

- We can refine our understanding of h via a bias-variance analysis:

Suppose f_X is smooth, so that we can Taylor expand around $x \in B_j$ as $f_X(y) \approx f_X(x) + (y-x) f_X'(x)$. This allows us to estimate the bias at x

as $B(x) = \mathbb{E}(\hat{f}_n(x)) - f_X(x)$

$$= \frac{1}{h}\int_{B_j} f_X(y)dy - f_X(x)$$

$$\approx \frac{1}{h}\int_{B_j} f_X(x) + (y-x) f_X'(x)dy - f_X(x)$$

$$= \frac{1}{h}\cdot f_X(x)\int_{B_j} dy + \frac{f_X'(x)}{h}\int_{B_j}(y-x)dy - f_X(x)$$

$$= \frac{f_X'(x)}{h}\left[\frac{1}{2}y^2 - xy\right]_{j-1}^{j}$$

$$= \frac{f'_{\tilde{X}}(x)}{h}\left[\frac{1}{2}h^2\cdot\left[j^2-(j-1)^2\right]-xh\right]$$

$$= f'_{\tilde{X}}(x)\cdot\left[\frac{1}{2}h\left[2\cdot j-1\right]-x\right]$$

$$= f'_{\tilde{X}}(x)\cdot\left[h\left[j-\frac{1}{2}\right]-x\right]$$

This holds for $x \in B_j$. So, we can compute the total bias-induced error as

$$\text{Bias}^2 = \int_0^1 B^2(x)\,dx = \sum_{j=1}^{N}\int_{B_j} B^2(x)\,dx$$

$$\approx \sum_{j=1}^{N}\int_{B_j}\left(f'_{\tilde{X}}(x)\cdot\left[h\left(j-\tfrac{1}{2}\right)-x\right]\right)^2 dx$$

$\tilde{x}_j \in B_j$ arbitrary

$$\approx \sum_{j=1}^{N}\left(f'_{\tilde{X}}(\tilde{x}_j)\right)^2\cdot\int_{B_j}\left[h\left(j-\tfrac{1}{2}\right)-x\right]^2 dx$$

$$= \sum_{j=1}^{N}\left(f'_{\tilde{X}}(\tilde{x}_j)\right)^2\cdot\frac{h^3}{12}$$

$$= \frac{h^2}{12}\underbrace{\sum_{j=1}^{N} h\cdot\left(f'_{\tilde{X}}(\hat{x}_j)\right)^2}_{\xrightarrow{h\to 0^+}\ \int_0^1\left(f'_{\tilde{X}}(x)\right)^2 dx}$$

$$\approx \frac{h^2}{12}\cdot\int_0^1\left(f'_{\tilde{X}}(x)\right)^2 dx.$$

- So, as expected, the bias improves as $h \to 0^+$. What about the variance? Well, we can estimate the variance at x as

$$V(x) = \frac{1}{h^2 \cdot n} \rho_j \cdot (1-\rho_j)$$

$$\approx \frac{1}{h^2 \cdot n} \cdot \rho_j \qquad \text{since } 1-\rho_j \approx 1 \text{ when } h \text{ is small}$$

Same argument for ρ_j as above $\rightarrow$

$$\approx \frac{1}{h^2 \cdot n} \cdot \left[h f_{\underline{x}}(x) + h f'_{\underline{x}}(x) \left(h\left(j - \tfrac{1}{2}\right) - x \right) \right]$$

$$= \frac{-f_{\underline{x}}(x)}{hn} + \text{Something} \to 0 \text{ as } h \to 0$$

$$\approx \frac{f_{\underline{x}}(x)}{hn}$$

$\Rightarrow$ Variance $= \int_0^1 V(x)\,dx$

$$\approx \int_0^1 \frac{f_{\underline{x}}(x)}{hn}\,dx$$

$$= \frac{1}{hn} \int_0^1 f_{\underline{x}}(x)\,dx$$

$$= \frac{1}{hn}.$$

- Hence, $MSE = \text{Bias}^2 + \text{Variance} \approx \underbrace{\frac{h^2}{12} \cdot \int_0^1 \left[f'_{\underline{x}}(x) \right]^2 dx}_{\text{want } h \text{ small}} + \underbrace{\frac{1}{hn}}_{\text{want } h \text{ large}}$!