# Lec 10: Introduction to Nonparametric Statistics

Yanjun Han

Nov. 19, 2024

__Nonparametric model:__ $Y \sim P_\theta$, $\theta$: infinite-dimensional, typically a function

(typically written as $Y \sim P_f$)

Canonical examples:

__Regression:__ given $(X_1, Y_1), \cdots, (X_n, Y_n) \sim P_{XY}$, estimate the regression function

$$f(x) := \mathbb{E}[Y \mid X = x]$$

__Density estimation:__ given $X_1, \cdots, X_n \sim f$ with an unknown density $f$, estimate $f$.

Other examples: in causal inference, interested in:

- causal function: $c(x) = \mathrm{Cov}(Y, W \mid X = x)$
- conditional/heterogeneous ATE (CATE):

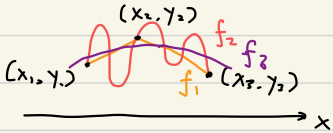
$$\tau(x) = \mathbb{E}[Y \mid W=1, X=x] - \mathbb{E}[Y \mid W=0, X=x]$$

Features of nonparametric models:

- model size > sample size; assumptions are necessary to prevent overfitting (typically smoothness or shape)

- MLE not well-defined / non-unique / hard to find

- explicit bias-variance tradeoff!




all are valid; but which is reasonable?

# Nonparametric regression

## A simple binning estimator

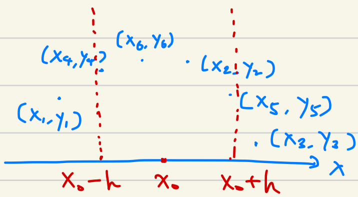
Assumption : $x_1, \cdots, x_n \in [0,1]$ nonrandom (fixed design)

Target: given $x_0 \in [0,1]$, estimate $f(x_0) = \mathbb{E}[Y \mid X = x_0]$

A simple estimator:

bandwidth parameter

$$\hat{f}_h(x_0) = \frac{\sum_{i : |x_i - x_0| \le h} Y_i}{\#\{i : |x_i - x_0| \le h\}}$$




( $h \to 0$ : overfit to the closest data point ;
$h \to \infty$ : underfit to the sample average of $y$ )

**Analysis** : assume that $f$ is $L$-Lipschitz, i.e. $|f'(x)| \le L \ \forall x$

(or equivalently, $|f(x) - f(y)| \le L|x - y| \ \forall x, y$ )

also, assume that $\mathrm{Var}(Y \mid X = x) \le \sigma_0^2$ for all $x$

### Variance of $\hat{f}_h(x_0)$ :

$$\mathrm{Var}(\hat{f}_h(x_0)) = \frac{\mathrm{Var}(\sum_{i : |x_i - x_0| \le h} Y_i)}{(\#\{i : |x_i - x_0| \le h\})^2} \quad \text{(non-random } \{x_i\}\text{)}$$

$$= \frac{\sum_{i : |x_i - x_0| \le h} \mathrm{Var}(Y_i)}{(\#\{i : |x_i - x_0| \le h\})^2} \quad \text{(independence)}$$

$$\le \frac{\sigma_0^2}{\#\{i : |x_i - x_0| \le h\}}$$

If $\{x_i\}$ are evenly spaced in $[0,1]$, then $\mathrm{Var}(\hat{f}_h(x_0)) = O\left(\frac{\sigma_0^2}{nh}\right)$.

## Bias of $\hat{f}_h(x_0)$:

$$|\mathrm{Bias}(\hat{f}_h(x_0))| = |\mathbb{E}[\hat{f}_h(x_0)] - f(x_0)|$$

$$= \left| \frac{\sum_{i: |x_i - x_0| \le h} f(x_i)}{\#\{i : |x_i - x_0| \le h\}} - f(x_0) \right| \quad (\mathbb{E}[y_i] = f(x_i))$$

$$= \left| \frac{\sum_{i: |x_i - x_0| \le h} (f(x_i) - f(x_0))}{\#\{i : |x_i - x_0| \le h\}} \right|$$

$$\le \frac{\sum_{i: |x_i - x_0| \le h} |f(x_i) - f(x_0)|}{\#\{i : |x_i - x_0| \le h\}} \quad \text{(triangle inequality)}$$

$$\le \frac{\sum_{i: |x_i - x_0| \le h} L|x_i - x_0|}{\#\{i : |x_i - x_0| \le h\}} \quad \text{(Lipschitz condition)}$$

$$\le Lh.$$

## Final mean-squared error (MSE)

$$\mathrm{MSE}(\hat{f}_h(x_0)) = \mathbb{E}(\hat{f}_h(x_0) - f(x_0))^2$$

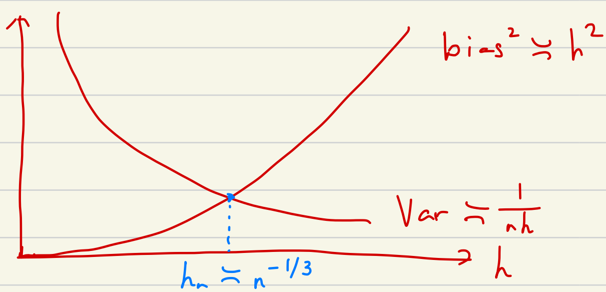
$$= \mathrm{Bias}(\hat{f}_h(x_0))^2 + \mathrm{Var}(\hat{f}_h(x_0))$$

$$= O\left(L^2 h^2 + \frac{\sigma_0^2}{nh}\right)$$

Optimal choice of $h$: $h = h_n = \left(\frac{\sigma_0^2}{nL^2}\right)^{1/3}$

Optimal MSE $= O(L^{2/3} \sigma_0^{4/3} n^{-2/3})$

Bias-variance tradeoff:



## Generalization: Nadaraya-Watson estimator

**Kernel**: a (non-negative) function $K: \mathbb{R}^d \to \mathbb{R}$ s.t. $\int_{\mathbb{R}^d} K(x)\,dx = 1$.

**Rescaled kernel**: for $h>0$, let $K_h(x) = \frac{1}{h^d} K(\frac{x}{h})$

Examples: rectangle/box kernel: $K(x) = \mathbb{1}(\|x\|_\infty \le \frac{1}{2})$

Gaussian kernel: $K(x) = \left(\frac{1}{2\pi}\right)^{d/2} \exp(-\frac{1}{2}\|x\|_2^2)$

Property: $\int_{\mathbb{R}^d} K_h(x)dx = \int_{\mathbb{R}^d} K_h(hz)\, h^d dz = \int_{\mathbb{R}^d} K(z)\,dz = 1$. $\forall h>0$.

**Kernel-regression estimator / Nadaraya-Watson estimator**

$$\hat{f}_h(x_0) = \frac{\sum_{i=1}^n K_h(x_0 - x_i)\, y_i}{\sum_{i=1}^n K_h(x_0 - x_i)}$$

- previous estimator corresponds to $K$ being the box kernel;
- an equivalent expression of $\hat{f}_h(x_0)$ is

$$\hat{f}_h(x_0) = \sum_{i=1}^n w_i(x_0)\, y_i, \quad \text{with} \quad w_i(x_0) = \frac{K_h(x_0 - x_i)}{\sum_{i=1}^n K_h(x_0 - x_i)}$$

being the "weight" of $x_i$ for $x_0$.

**Analysis**. Assume that: 1) $|K(x)| \le B$ for all $x$;

2) $K(x) = 0$ for all $|x| \ge M$.

$$\mathrm{Var}(\hat{f}_h(x_0)) \le \frac{\sum_{i=1}^n K_h^2(x_0 - x_i)\,\sigma_0^2}{\left(\sum_{i=1}^n K_h(x_0 - x_i)\right)^2} \le \frac{\frac{B}{h^d}\,\sigma_0^2}{\sum_{i=1}^n K_h(x_0 - x_i)}$$

$$= \frac{B\sigma_0^2}{nh^d} \cdot \frac{1}{\underbrace{\frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right)}_{\text{expect to be a constant}}}$$

$$|\mathrm{Bias}(\hat{f}_h(x_0))| = \left| \frac{\sum_{i=1}^n K_h(x_0 - x_i) f(x_i)}{\sum_{i=1}^n K_h(x_0 - x_i)} - f(x_0) \right|$$

$$\leq \frac{\sum_{i=1}^n K_h(x_0 - x_i)\, |f(x_i) - f(x_0)|}{\sum_{i=1}^n K_h(x_0 - x_i)}$$

$K_h(x_0 - x_i) = 0$ if $|x_0 - x_i| \geq hM$

$\leq L|x_i - x_0|$

$$\leq \frac{\sum_{i=1}^n K_h(x_0 - x_i) \cdot LhM}{\sum_{i=1}^n K_h(x_0 - x_i)} = LMh$$

$$\mathrm{MSE}(\hat{f}_h(x_0)) = O\left(L^2 M^2 h^2 + \frac{B\sigma_0^2}{nh^d}\right) = O\left(n^{-\frac{2}{2+d}}\right)$$

optimal bandwidth $h = h_n \asymp n^{-\frac{1}{2+d}}$

Capturing higher smoothness of $f$: next lecture (local polynomials / splines)

Density estimation: estimate $f$ from $X_1, \dots, X_n \sim f$.

Kernels are still useful: let $K$ be a kernel with $\int_{\mathbb{R}^d} x K(x)\, dx = 0$.

Kernel density estimator (KDE):

$$\hat{f}_h(x_0) = \frac{1}{n} \sum_{i=1}^n K_h(x_0 - X_i)$$

- Intuition: when $K$ is the box kernel,

$$\hat{f}_h(x_0) = \frac{\#\{i : x_i \text{ lies in the box centered at } x_0 \text{ of edgelength } h\}}{nh^d}$$

$$\approx \frac{f(x_0) \cdot nh^d}{nh^d} = f(x_0)$$

Analysis: assume that $\|f''\|_\infty \leq L$, $\int x^2 K(x)\, dx < \infty$, $\int K^2(x)\, dx < \infty$, and $d = 1$.

$$\mathrm{Var}(\hat{f}_h(x_0)) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(K_h(x_0 - X_i))$$
$$\leq \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}[K_h(x_0 - X_i)^2]$$
$$= \frac{1}{nh^2}\int K\left(\frac{x_0 - x}{h}\right)^2 dx$$
$$= \frac{1}{nh}\int K^2(z)\,dz \qquad (x := x_0 - hz)$$

$$|\mathrm{Bias}(\hat{f}_h(x_0))| = |\mathbb{E}[K_h(x_0 - X_1)] - f(x_0)|$$
$$= \Bigg| \underbrace{\int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) f(x)\,dx}_{\text{convolution of } f \text{ and } K_h} - f(x_0)\Bigg|$$
$$= \left|\int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right)(f(x) - f(x_0))\,dx\right| \qquad \left(\int K_h(x)\,dx = 1\right)$$
$$= \left|\int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right)\Big(f(x) - f(x_0) - f'(x_0)(x - x_0)\Big)\,dx\right|$$
$$\left(\int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right)(x_0 - x)\,dx = h \cdot \int z K(z)\,dz = 0\right)$$
$$\leq \int \frac{1}{h} K\left(\frac{x_0 - x}{h}\right) \cdot \frac{L}{2}(x_0 - x)^2\,dx$$
$$= \frac{Lh^2}{2}\int z^2 K(z)\,dz \qquad (x = x_0 - hz)$$

$$\mathrm{MSE}(\hat{f}_h(x_0)) = O\left(L^2 h^4 + \frac{1}{nh}\right) \underset{\uparrow}{=} O(n^{-4/5})$$

optimal bandwidth $h = h_n \asymp n^{-1/5}$

View Nadaraya-Watson as KDE:

$$\mathbb{E}[Y \mid X=x] = \frac{\int y f(x,y)\,dy}{\int f(x,y)\,dy} \approx \frac{\int y \cdot \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)K_h(y-Y_i)\,dy}{\int \frac{1}{n}\sum_{i=1}^{n} K_h(x-X_i)K_h(y-Y_i)\,dy}$$

$$= \frac{\sum_{i=1}^{n} K_h(x-X_i)\,Y_i}{\sum_{i=1}^{n} K_h(x-X_i)}$$

## Nearest-neighbor density estimator

Define $r_i = \|X_i - x_0\|_2$ as the distance between $X_i$ and $x_0$

For $k=1,2,\cdots,n$, let $r_{(k)}$ be the $k$-th smallest element of $(r_1,\cdots,r_n)$

($k$-th nearest neighbor)

$$\boxed{\hat{f}_k(x_0) = \frac{k/n}{\mathrm{Vol}_d(r_{(k)})} = \frac{k/n}{\frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}\, r_{(k)}^d}}$$

volume of $d$-dim ball of radius $r_{(k)}$

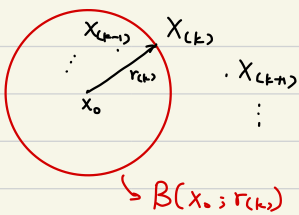
Intuition: $f(x_0)\cdot \mathrm{Vol}_d(r_{(k)})$

$\approx \int_{B(x_0; r_{(k)})} f(x)\,dx$

$=$ actual prob. of $X \in B(x_0; r_{(k)})$

$\approx$ empirical prob. of $X \in B(x_0; r_{(k)})$

$= \frac{k}{n}$.




Rigorous claim: $\int_{B(x_0; r_{(k)})} f(x)\,dx \sim \mathrm{Beta}(k, n-k+1)$

Pf. LHS $=$ $k$-th smallest element of $\left\{ Z_i = \int_{B(x_0; \|X_i - x_0\|_2)} f(x)\,dx \right\}_{i=1}^{n}$

where $\mathbb{P}(Z_i \le t) = \mathbb{P}(\|X_i - x_0\|_2 \le g^{-1}(t))$ $\left(g(r) := \int_{B(x_0;r)} f(x)\,dx\right)$

$= g(g^{-1}(t)) = t \Rightarrow Z_i \sim \mathrm{Unif}[0,1]$ $\square$

Analysis (Optional) Assume that $\|f''\|_\infty \le L$ & $c \le f(x) \le C$ for all $x \in \mathrm{supp}(f)$.

Step I.
$$\left| \int_{B(x_0; r_{(k)})} f(x)\,dx - f(x_0)\cdot \mathrm{Vol}_d(r_{(k)}) \right|$$
$$= \left| \int_{B(x_0; r_{(k)})} (f(x) - f(x_0))\,dx \right|$$
$$= \left| \int_{B(x_0; r_{(k)})} [f(x) - f(x_0) - f'(x_0)(x - x_0)]\,dx \right|$$
$$\le \int_{B(x_0; r_{(k)})} \frac{L}{2} \|x - x_0\|_2^2\,dx \le \frac{L}{2} r_{(k)}^2 \mathrm{Vol}_d(r_{(k)})$$

Step II.
$$\mathbb{E} \left| \int_{B(x_0; r_{(k)})} f(x)\,dx - \hat{f}_k(x_0)\cdot \mathrm{Vol}_d(r_{(k)}) \right|^2$$
$$= \mathbb{E} \left| \mathrm{Beta}(k, n+1-k) - \frac{k}{n} \right|^2 = O\left(\frac{k}{n^2}\right).$$

Step III. Since $f(x) \asymp 1$ everywhere,
$$\frac{k}{n} \overset{\text{whp}}{\asymp} \int_{B(x_0; r_{(k)})} f(x)\,dx \asymp \mathrm{Vol}_d(r_{(k)})$$
$$\Longrightarrow \mathrm{Vol}_d(r_{(k)}) \asymp \frac{k}{n} \Longrightarrow r_{(k)} \asymp \left(\frac{k}{n}\right)^{1/d}.$$

Conclusion.
$$f(x_0)\,\mathrm{Vol}_d(r_{(k)}) \overset{O(r_{(k)}^2 \mathrm{Vol}_d(r_{(k)}))}{\approx} \int_{B(x_0; r_{(k)})} f(x)\,dx$$
$$\overset{O(\sqrt{k}/n)}{\approx} \hat{f}_k(x_0)\cdot \mathrm{Vol}_d(r_{(k)})$$
$$\Longrightarrow \mathrm{MSE}(\hat{f}_k(x_0)) = O\left( r_{(k)}^4 + \frac{k/n^2}{\mathrm{Vol}_d(r_{(k)})^2} \right)$$
$$= O\left( \left(\frac{k}{n}\right)^{4/d} + \frac{1}{k} \right) \underset{k = k_n \asymp n^{\frac{4}{4+d}}}{=} O\left(n^{-\frac{4}{4+d}}\right)$$

(matching the KDE result for $d=1$)