Causal Machine Learning – Fall 2023

Week 4: Nonparametrics

Max H. Farrell & Sanjog Misra

Topics to cover

1. Crash Course in Nonparametrics

Definitions

What does it mean for a model to be

1. Parametric
2. Nonparametric
3. Semiparametric

Examples:

- OLS
- ML
- ?

Motivation: flexibility in first stage

Last class:

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\mathbb{E}\left[Y \mid T=1, X\right]\right] \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(x_i),\ \hat{\mu}_1(x_i) = x_i'\hat{\beta}_1$$

- Today: Make $\hat{\mu}_1(x_i)$ more flexible
- Later: How that messes up the second step

Approximating a smooth function

- Smoothness means that $f(x)$ "can't change too fast"
- Fundamental idea underpinning most (all?) nonparametrics.
- The function $f(x) : \Re \to \Re$ is smooth means

$$x_1 \approx x_2 \quad \Rightarrow \quad f(x_1) \approx f(x_2)$$

- Formalize with a Taylor approximation. If $f(x) : \Re \to \Re$ has p derivatives, then:

$$\begin{aligned} f(x_2) = f(x_1) + f'(x_1)(x_2 - x_1) + f^{(2)}(x_1)(x_2 - x_1)^2/2 \\ + \cdots + f^{(p)}(x_1)(x_2 - x_1)^p/p! \end{aligned}$$

- Multivariate version is the same, just more notation.

Approximating a smooth function

How bad can a linear regression approximation be?

- Imagine $Y = f(X) + \varepsilon$ and you fit $Y = \beta_0 + \beta_1 X + \varepsilon$
- The true function is

$$\begin{aligned} f(x) &= f(0) + f'(0)(x-0) + f^{(2)}(0)(x-0)^2/2 + \cdots \\ &= f(0) + f'(0)x + \frac{f^{(2)}(0)}{2}x^2 + \cdots \\ &=: \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \end{aligned}$$

- Estimate $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

Approximating a smooth function

Errors in $f(x)$ versus $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

1. $\hat{\beta}_j \neq \beta_j$
   - We already know that $\hat{\beta} - \beta = O_p(1/\sqrt{n})$. Why?
2. $f(x) \neq \beta_0 + \beta_1 x$
   - Because we left off all those other derivatives
   - Because x is not 0

Approximating a smooth function

Tuning parameter choice

- $\hat{f}(x) - f(x) = \sqrt{\frac{J}{n}} + J^{-1}$
- Bigger J (smaller bins) $\rightarrow$ better approximation, higher variance
- Bigger polynomial $\rightarrow$ better approximation, higher variance
- $J^{\star} \asymp n^{1/3}$
- Optimal for estimating $f \neq$ optimal for two-step

Curse of dimensionality

- For $x \in \mathbb{R}^d$:

$$\hat{f}(x) - f(x) = \sqrt{\frac{J^d}{n}} + J^{-1}$$

- Exponentially worse!

Approximating a smooth function

Other classical nonparametric estimators are very similar

- Polynomial K in J bins: $\sqrt{\frac{J^d}{n}} + J^{-K-1}$
- Kernelsof order P: $\frac{1}{\sqrt{nh^d}} + h^P$
- Series: $\sqrt{\frac{K}{n}} + K^{-\alpha}$
- In general:

$$\text{Var} = \frac{1}{\text{effective sample size}} = \frac{\#\text{ params}}{n}$$

$$\text{Bias} = (\#\text{ params})^{-(\text{smoothness})}$$

High Dimensional Models

What to do if $d = \dim(X)$ is "large"?

- Asymptotics: d fixed, $d \to \infty$, $d/n \to ?$
- Generic version: $f(x) = f_n(x) +$ bias
- NP version: local/simple + bias
- High-Dim version: need a functional form assumption for $f_n(x)$
- Lasso: $f_n(x) = x'\boldsymbol{\beta}$, for $\boldsymbol{\beta} \in \mathbb{R}^d$, $\|\boldsymbol{\beta}\|_0 = s = o(n)$
- rate: $\dfrac{\#\text{ params}}{n} = \dfrac{s}{n}$.
- Don't know **which** s terms, search cost $= \frac{s \log(d)}{n}$.