Lecture 11

在 density estimation 中，我们的思路为：Histogram estimator $\longrightarrow$ Rosenblatt estimator $\longrightarrow$ Kernel estimator，类似的思路也可以应用于 nonparametric regression 中.

§1 Nonparametric regression

1. Definition: Nonparametric regression 的 settings

Nonparametric regression 的 settings 为：

$$Y_i = m(X_i) + \varepsilon_i \quad , \quad i = 1, \cdots, n$$

其中，$\{\varepsilon_i\}_{1\le i\le n}$ 为 i.i.d. r.v. with mean 0 and variance σ^2

注：Linear regression model 是 nonparametric regression 的特例：

where $\{\epsilon_i\}_{1\le i\le n}$ are i.i.d random variables with mean 0 and vairance σ^2. It's a nonparametric statistical problem cause the unknown parameter of interests, i.e., the function $m(\cdot)$ is of infinite dimension. When $m(\cdot)$ is a linear function, i.e.,

$$Y_i = X_i\beta + \epsilon_i, \quad i = 1, \cdots, n,$$

then we simplified our problem to the more commonly seen linear regression model, where estimating $m(\cdot)$ is equivalent to estimate β, and the least square estimator is given by $\hat{\beta} = (X^TX)^{-1}X^TY$. However, when $m(\cdot)$ is beyond linear function, the estimation gets tricky.

2. 一个 intuitive estimator

一个 intuitive estimator 为

$$\hat{m}(x) = \underbrace{\frac{\sum_{i=1}^n Y_i \cdot 1(X_i = x)}{\sum_{i=1}^n 1(X_i = x)}}_{\text{所有}X=x\text{的样本的}Y\text{的均值}} = m(x) + \underbrace{\frac{\sum_{i=1}^n \varepsilon_i \cdot 1(X_i = x)}{\sum_{i=1}^n 1(X_i = x)}}_{\text{所有}X=x\text{的样本的}\varepsilon\text{的均值}}$$

注：缺点：x 有 infinite 个时，$X_i = x$ 的个数会非常少

类似于 density estimation，我们希望 x 附近的样本都能有贡献.

3. Definition: Rosenblatt estimator

若 nonparametric model 下的 independent sample 为 $\{(Y_i, X_i)\}_{1\le i\le n}$，

则 Rosenblatt estimator of $m(x)$ 为：

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i \cdot 1(x - \frac{h}{2} \le X_i \le x + \frac{h}{2})}{\sum_{i=1}^n 1(x - \frac{h}{2} \le X_i \le x + \frac{h}{2})}$$

4. Definition: Nadaraya-Watson kernel estimator

若 nonparametric model 下的 independent sample 为 $\{(Y_i, X_i)\}_{1\le i\le n}$，

则 Nadaraya-Watson kernel estimator of $m(x)$ 为：

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i \cdot K(\frac{x - X_i}{h})}{\sum_{i=1}^n K(\frac{x - X_i}{h})} \quad \text{(always positive)}$$

5.* Theorem: Nadaraya-Watson kernel estimator 的 L_1 consistency

*** Theorem 5.3** (♣ **L_1 consistency of the Nadaraya-Watson Kernel Estimator**). *Assume the symmetric kernel function $K(\cdot)$ is bounded and having a compact support, further, we have $\{(X_i, \epsilon_i)\}_{1 \le i \le n}$ being a random sample with X_0 being an independent copy of the X_i. Assume*

$$\lim_{n \to \infty} h_n = 0, \quad \lim_{n \to \infty} nh_n = \infty.$$

then $\mathbb{E}|\hat{m}(X_0) - m(X_0)| \to 0$.

6.* Theorem: Nadaraya-Watson kernel estimator 的 bandwidth 的选取

*** Theorem 5.4** (♣ **Bandwidth Selection in the Nadaraya-Watson Kernel Estimator**). *Assume we have a random sample $\{(Y_i, X_i)\}_{1 \le i \le n}$ where*

$$Y_i = m(X_i) + \sigma(X_i) \cdot \epsilon_i, \quad i = 1, \cdot, n$$

where $\{X_i\}_{1 \le i \le n}$ are i.i.d with density function f_X and $\{\epsilon_i\}_{1 \le i \le n}$ are i.i.d with mean 0 and variance 1. Suppose $\{X_i\}_{1 \le i \le n}$ are mutually independent with $\{\epsilon_i\}_{1 \le i \le n}$. For the Nadaraya-Watson estimator $\hat{m}(x)$ of estimating this $m(x)$, assume our symmetric kernel function $K(\cdot)$ is bounded and having a compact support and finite moment generating function, with

$$\lim_{n \to \infty} h_n = 0, \quad \lim_{n \to \infty} nh_n = \infty.$$

then

$$\boxed{\begin{aligned} \mathrm{Bias}(\hat{m}(x), m(x)) &\to \left(\frac{1}{2}m''(x) + \frac{m'(x) f_X'(x)}{f_X(x)}\right)\left(\int \mu^2 K(\mu) d\mu\right) h_n^2 \\ \mathrm{Var}\,\hat{m}(x) &\to \frac{\sigma^2(x)}{f_X(x) n h_n} \int K^2(\mu) d\mu \end{aligned}}$$

and the optimal bandwidth h_n is again given by

$$\boxed{\begin{aligned} h_n = &\left[\int \mu^2 K(\mu) d\mu\right]^{-2/5} \left[\int K^2(y) dy\right]^{1/5} \left[\int \frac{\sigma^2(x)}{f_X(x)} dx\right]^{1/5} \\ &\times \left[\int \left(\frac{1}{2}m''(x) + \frac{m'(x) f_X'(x)}{f_X(x)}\right) dx\right]^{-1/5} n^{-1/5}. \end{aligned}}$$

convergence rate 为 $n^{-4/5}$

In practice, one may again use rule of thumb or slove-the-equation procedure in order to have the optimal bandwidth h_n.

§2 Nonparametric regression 的 applications

例 1: (估计 conditional second moment)

• *Example* 6.1 (**Estimating the Conditional Second Moment**). Assume we have a random sample $\{(Y_i, X_i)\}_{1 \le i \le n}$, where each $Y_i = (Y_{i1}, \cdots, Y_{ip})^T \in \mathcal{M}_{p \times 1}$. Say we are interested in estimating the second moment of Y_i condition on $X_i = x$, for instance, we are interested in

$$m_{1,j}(x) = \mathbb{E}(Y_{ij} | X_i = x)$$

$$m_{2,j}(x) = \mathrm{Var}(Y_{ij} | X_i = x) = \mathbb{E}(Y_{ij}^2 | X_i = x) - \Big[\mathbb{E}(Y_{ij} | X_i = x)\Big]^2$$

$$\begin{aligned} m_{3,jk}(x) &= \mathrm{Cov}(Y_{ij}, Y_{ik} | X_i = x) \\ &= \mathbb{E}(Y_{ij} Y_{ik} | X_i = x) - \Big[\mathbb{E}(Y_{ij} | X_i = x)\Big]\Big[\mathbb{E}(Y_{ik} | X_i = x)\Big]. \end{aligned}$$

Please give an estimator for each of the above terms.

Answer. A reasonable choice is to fit all

$$\mathbb{E}(Y_{ij}|X_i = x), \quad \mathbb{E}(Y_{ij}^2|X_i = x), \quad \mathbb{E}(Y_{ij}Y_{ik}|X_i = x)$$

to nonparametric regression models separately. Therefore, the corresponding Nadaraya-Watson Kernel estimator for each of them is given by

$$\boxed{\begin{aligned}
\hat{\mathbb{E}}(Y_{ij}|X_i = x) &= \frac{\sum_{i=1}^n Y_{ij} K\left((x - X_i)/h_1\right)}{\sum_{i=1}^n K\left((x - X_i)/h_1\right)}, \\
\hat{\mathbb{E}}(Y_{ij}^2|X_i = x) &= \frac{\sum_{i=1}^n Y_{ij}^2 K\left((x - X_i)/h_2\right)}{\sum_{i=1}^n K\left((x - X_i)/h_2\right)}, \\
\hat{\mathbb{E}}(Y_{ij}Y_{ik}|X_i = x) &= \frac{\sum_{i=1}^n Y_{ij}Y_{ik} K\left((x - X_i)/h_3\right)}{\sum_{i=1}^n K\left((x - X_i)/h_3\right)},
\end{aligned}}$$

and correspondingly, by plug-in the above estimator, we obtain

$$\boxed{\begin{aligned}
\hat{m}_{1,j}(x) &= \frac{\sum_{i=1}^n Y_{ij} K\left((x - X_i)/h_1\right)}{\sum_{i=1}^n K\left((x - X_i)/h_1\right)}, \\
\hat{m}_{2,j}(x) &= \frac{\sum_{i=1}^n Y_{ij}^2 K\left((x - X_i)/h_2\right)}{\sum_{i=1}^n K\left((x - X_i)/h_2\right)} - \left[\frac{\sum_{i=1}^n Y_{ij} K\left((x - X_i)/h_1\right)}{\sum_{i=1}^n K\left((x - X_i)/h_1\right)}\right]^2 \\
\hat{m}_{3,jk}(x) &= \frac{\sum_{i=1}^n Y_{ij}Y_{ik} K\left((x - X_i)/h_3\right)}{\sum_{i=1}^n K\left((x - X_i)/h_3\right)} \\
&\quad - \left[\frac{\sum_{i=1}^n Y_{ij} K\left((x - X_i)/h_1\right)}{\sum_{i=1}^n K\left((x - X_i)/h_1\right)}\right]\left[\frac{\sum_{i=1}^n Y_{ik} K\left((x - X_i)/h_1\right)}{\sum_{i=1}^n K\left((x - X_i)/h_1\right)}\right].
\end{aligned}}$$

Notice here we have used three different bandwidth h_1, h_2 and h_3, which is necessary in order to have a optimal convergence rate of the MISE, and each of them may be obtained using the slove-the-equation bandwidth estimator. □

例2：（估计 conditional p-th quantile）

• *Example* 6.2 (**Estimating the Conditional p−th Quantile**). Assume we have a random sample $\{(Y_i, X_i)\}_{1\le i\le n}$, and we are interested in estimating the conditional $p-$th quantile of Y_i condition on $X_i = x$, i.e.,

$$\xi_p = \inf_{\xi} \left\{\xi : \mathbb{P}(Y_i \le \xi|X_i = x) \ge p\right\}$$

Please give an estimator for ξ_p.

Answer. Estimating ξ_p directly appears to be a difficulty problem. Therefore, we may obtain our estimator $\hat{\xi}_p$ of ξ_p from an estimated condition distribution function, i.e., define

$$\hat{\xi}_p = \inf_{\xi} \left\{\xi : \hat{F}_{Y_i}(\xi|X_i = x) \ge p\right\}.$$

Thus, we switch our task from estimating ξ_p directly to obtain an estimator $\hat{F}_{Y_i}(\xi|X_i = x)$ of $\mathbb{P}(Y_i \le \xi|X_i = x)$. Notice that

$$\boxed{\mathbb{P}(Y_i \le \xi|X_i = x) = \mathbb{E}\Big[\mathbb{1}(Y_i \le \xi)|X_i = x\Big]}$$

Therefore, for each given xi and x, we may use the Nadaraya-Watson kernel estimator, given is

$$\boxed{\hat{F}_{Y_i}(\xi|X_i = x) = \frac{\sum_{i=1}^n \mathbb{1}(Y_i \le \xi) K\left((x - X_i)/h\right)}{\sum_{i=1}^n K\left((x - X_i)/h\right)}.}$$

By plug-in this $\hat{F}$ we obtained our estimator for the conditional $p-$th quantile. □