A short notes on Nonparametric regression

Tong Zhou*

October 31, 2020

This notes is mainly based on the following materials:

- All of Nonparametric Statistics, by *Larry Wasserman*
- A Primer on Regression Splines, by *Jeffrey S. Racine*
- Lecture 3: Regression: Nonparametric Approaches, STAT 535 SML Autumn 2019, by *Yen-chi Chen*
- Lecture notes: Nonparametric Regression and Classification, Statistical Machine Learning, Spring 2017, by *Ryan Tibshirani and Larry Wasserman*

Given pairs of data $(x_1, Y_1), \cdots, (x_n, Y_n)$, in which Y is called **response variable** and x is called **covariate**, our goal is to relate them by:

$$Y_i = r(x_i) + \varepsilon_i, \quad \mathbf{E}[\varepsilon_i] = 0, i = 1, \cdots, n.$$

where $r()$ is called the **regression function**. In the machine learning community, the goal is to "learn" r under weak assumptions. The estimator of r is denoted by $\widehat{r}_n(x)$, which is often referred to as a **smoother**.

Like the traditional statistics, there are two perspectives regarding the covariate x. First, x can be assumed to be *fixed*. This assumption does not hurt much, only in the purpose of simplifying derivations and easing treatment of notations. Once x is assumed be some realizations of random variables, all relevant notions can be augmented by using a "conditional" operator. For example, the regression function $r(x)$ can be written as the condition mean of Y on $X = x$, i.e.

$$r(x) = \mathbf{E}[Y|X = x].$$

We have mentioned r is called a smoother. Before plunging into the formal treatment of nonparametric regression, we first define the notion of **linear smoother**.

*Johns Hopkins University, tzhou11@jhu.edu

Definition 1. *An estimator $\widehat{r}_n$ of r is a* ***linear smoother*** *if, for each x, there exists a vector $\ell(x) = (\ell_1(x), \cdots, \ell_n(x))^\top$ such that*

$$\widehat{r}_n(x) = \sum_{i=1}^{n} \ell_i(x) Y_i.$$

Some related quantities are:

- **fitted values:** $\boldsymbol{r} = (\widehat{r}_n(x_1), \cdots, \widehat{r}_n(x_n))^\top$
- **Smoothing matrix / hat matrix:** $L = (\ell(x_1), \cdots, \ell(x_n))^\top$ and $L_{ij} = \ell_j(x_i)$.
- **effective kernel** for $r(x_i)$: $\ell(x_i)^\top$ the i-th row of L with $\sum_{i=1}^{n} \ell_i(x) = 1$.
- **effective degrees of freedom:** $\nu = \operatorname{tr}(L)$.
- $\boldsymbol{r} = LY$.

NOTE

For the regression model:

$$Y_i = r(x_i) + \varepsilon_i$$

The estimator $\widehat{r}_n$ of r is to minimize the sum of squares:

$$\sum_{i=1}^{n} (Y_i - \widehat{r}_n(x_i))^2 .$$

There are two extreme cases:

1. Minimizing over *ALL linear functions* $\Longrightarrow$ **least square estimators**.
2. Minimizing over *ALL functions* $\Longrightarrow$ **a function that interpolates the data**.

There are several solutions:

1. Local regression: use a locally weighted sums of squares.

$$\sum_{i=1}^{n} w_i(x) (Y_i - a)^2 .$$

2. Penalized sums of squares :

$$M(\lambda) = \sum_{i} (Y_i - \widehat{r}_n(x_i))^2 + \lambda J(r)$$

Example 1 (Local Regression). *Suppose $x_i \in \mathbb{R}$ is scalar, and consider the regression model $Y_i = r(x_i) + \varepsilon_i$. Idea of estimating $r(x)$ is by taking a weighted average of Y_i, giving higher weight to those points near x.*

- ***Nadaraya-Watson kernel estimator:***

$$\widehat{r}_n(x) = \sum_{i=1}^{n} \ell_i(x) Y_i.$$

- *K is a kernel and*

$$\ell_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

- *Example 5.26 (boxcar kernel:) Suppose that $x_i \in [a,b], i = 1, \cdots, n$. Divide (a,b) into m equally spaced bins denoted by $B_1, B_2, \cdots, B_m$. Define $\widehat{r}_n(x)$ by*

$$\widehat{r}_n(x) = \frac{1}{k_j} \sum_{i:\, x_i \in B_j} Y_i, \ \ \textit{for} \ \ x \in B_j$$

where k_j is the number of points in B_j. So for $x \in B_j$, $\ell_i(x) = \frac{1}{k_j}$ if $x_i \in B_j$ and $\ell_i = 0$ otherwise.

Example 2 (Penalized Regression). *Goal: to minimize the **penalized sums of squares:***

$$M(\lambda) = \sum_i (Y_i - \widehat{r}_n(x_i))^2 + \lambda J(r)$$

*where $J(r)$ is some **roughness penalty**, λ controls the amount of smoothing.*

1. *$\lambda \to 0$: the interpolating function.*
2. *$\lambda \to \infty$: the least squares line.*
3. *$0 < \lambda < \infty$: use **splines,** a special piecewise polynomial.*

A special case on the penalty term is such that

$$J(r) = \int |r''(x)|^2 \, \mathrm{d}x \tag{1}$$

Before diving into details, we first look at an important theorem:

Theorem 1. *The function $\widehat{r}_n(x)$ that minimizes $M(\lambda)$ with penalty defined in 1 is a natural cubic spline with knots at the data points. The estimator $\widehat{r}_n$ is called a **smoothing spline**.*

Now let us get back to some notions underlined in this theorem:

- Let $\xi_1 < \xi_2 < \cdots < \xi_p$ be set of ordered points-called **knots**.
- **cubic spline:** is a continuous function r such that (i) r is a cubic piecewise polynomial over each bins $(\xi_1, \xi_2), \cdots$ and (ii) r has continuous first and second derivatives at the knots.

However, Theorem 1 only asserts that the minimizer of the penalty regression problems exists and belongs to the class of cubic splines. It does not provide guidance on what the cubic spline looks like. The next theorem fills a gap:

Theorem 2. *Let $\xi_1 < \xi_2 < \cdots < \xi_p$ be knots contained in an interval (a, b). Define $h_1(x) = 1, h_2(x) = x, h_3(x) = x^2, h_4(x) = x^3, h_{j+4}(x) = (x - \xi_j)_+^3$ for $j = 1, \cdots, p$. The functions $\{h_1, \cdots, h_{k+4}\}$ form a basis for the set of cubic splines at these knots, called the **truncated power basis**. Thus, any cubic spline $r(x)$ with these knots can be written as*

$$r(x) = \sum_{j=1}^{k+4} \beta_j h_j(x).$$

Remark 1.
- The space of the 3-order splines with knots $\xi_1, \cdots, \xi_p$ has dimension $p + 4$.
- The basis of functions is called **truncated power basis**.
- More general, the k-order spline r is a piecewise polynomial function of degree k that is continuous and has continuous derivatives of orders $1, \cdots, k-1$, at its knot points. Specifically, its corresponding *truncated power basis* is $g_1, g_2, \cdots, g_{p+k+1}$ such that

$$g_1(x) = 1, g_2(x) = x, \cdots, g_{k+1}(x) = x^k,$$
$$g_{k+1+j}(x) = (x - \xi_j)_+^k, j = 1, \cdots, p.$$

  and its dimension is $p + k + 1$.

To gain more computational edge, another more popular spline is **B-spline**, where its basis functions are defined recursively.

Let $\xi_0 = a$ and $\xi_{p+1} = b$. Define the appended $2M$ knots by:

$$\tau_1 \leqslant \tau_2 \leqslant \tau_3 \leqslant \cdots \leqslant \tau_M \leqslant \xi_0,$$

and

$$\xi_{p+1} \leqslant \tau_{k+M+1} \leqslant \cdots \leqslant \tau_{p+2M}$$

This appending is needed due to the recursive nature of the B-spline. Now for the j-th order of spline, the basis functions can be defined as:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } \tau_i \leqslant x \leqslant \tau_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

for each i. Next, for $m \leqslant M$, we define

$$B_{i,m}(x) = \alpha_{i,m-1} B_{i,m-1}(x) + (1 - \alpha_{i+1,m}) B_{i+1,m}(x)$$

where

$$\alpha_{i,m-1}(x) = \begin{cases} \frac{x-\tau_i}{\tau_{i+m-1}-\tau_i}, & \text{if } \tau_{i+m-1} \neq \tau_i \\ 0, & \text{otherwise} \end{cases}$$

Now all relevant notations are squared away, the following theorem dictates the role of B-spline:

Theorem 3. *The function $\{B_{i,4}, i = 1, \cdots, k+4\}$ are a basis for the set of cubic splines. The are called the* ***B-spline basis functions****.*

Let's switch back to the original setting: $(x_1, Y_1), \cdots, (x_n, Y_n)$. Then $\widehat{r}_n$ is a natural cubic spline:

$$\widehat{r}_n(x) = \sum_{j=1}^{N} \widehat{\beta}_j B_j(x)$$

where $N = n + 4$ and $\widehat{\beta} = \left(\widehat{\beta}_1, \cdots, \widehat{\beta}_N\right)^\top$.

The minimization problem then can be expressed as

Minimization of Penalized Regression Estimator

$$\text{minimize: } \left(Y - B\beta\right)^\top \left(Y - B\beta\right) + \lambda \beta^\top \Omega \beta$$

where $B_{ij} = B_j(X_i)$ and $\Omega_{jk} = \int B_j''(x) B_k''(x)\, \mathrm{d}x$

The value of β that minimizes the problem is:

$$\widehat{\beta} = \left(B^\top B + \lambda \Omega\right)^{-1} B^\top Y.$$