Lecture Notes II.1 – Bias and variance in Kernel Regression

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

October, 2021

An elementary analysis

Bias, Variance and h for $x \in \mathbb{R}$

Kernel regression by Nadaraya-Watson

$$\hat{y}(x) \;=\; \frac{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right)} \tag{1}$$

Let $w_i = \frac{b\left(\frac{||x-x^i||}{h}\right)}{\sum_{i'=1}^{n} b\left(\frac{||x-x^{i'}||}{h}\right)}$.

Assumptions

A0 For simplicity, in this analysis we assume $x \in \mathbb{R}$.

A1 There is a true smooth[1] function $f(x)$ so that

$$y \;=\; f(x) + \varepsilon, \tag{2}$$

where ε is sampled independently for each x from a distribution P_ε, with $E_{P_\varepsilon}[\varepsilon] = 0$, $Var_{P_\varepsilon}(\varepsilon) = \sigma^2$.

A2 The kernel $b(z)$ is smooth, $\int_{\mathbb{R}} b(z)dz = 1$, $\int_{\mathbb{R}} zb(z) = 0$, and we denote $\sigma_b^2 = \int_{\mathbb{R}} z^2 b(z)dz$, $\gamma_b^2 = \int_{\mathbb{R}} b^2(z)dz$.

In this first analysis, we consider that the values x, $x^{1:N}$ are fixed; hence, the randomness is only in $\varepsilon^{1:N}$.

[1] with continuous derivatives up to order 2

Expectation of $\hat{y}(x)$ – a simple analysis

Expanding f in Taylor series around x we obtain

$$f(x^i) \;=\; f(x) + f'(x)(x^i - x) + \frac{f''(x)}{2}(x^i - x)^2 + o((x^i - x)^2) \tag{3}$$

We also have

$$y^i = f(x^i) + \varepsilon^i. \tag{4}$$

We now write the expectation of $\hat{y}(x)$ from (1), replacing in it y^i and $f(x^i)$ as above. What we would like to happen is that this expectation equals $f(x)$. Let us see if this is the case.

$$E_{P_\varepsilon^n}[\hat{y}(x)] \;=\; E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i y^i\right] \;=\; E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \left(f(x^i) + \varepsilon^i\right)\right] \tag{5}$$

$$= \sum_{i=1}^n w_i f(x) + \sum_{i=1}^n w_i f'(x)(x^i - x) + \sum_{i=1}^n w_i \frac{f''(x)}{2}(x^i - x)^2 + \underbrace{E_{P_\varepsilon^n}\left[\sum_{i=1}^n w_i \varepsilon^i\right]}_{=0]} \tag{6}$$

$$= \underbrace{f(x) + f'(x)\sum_{i=1}^n w_i(x^i - x) + \frac{f''(x)}{2}\sum_{i=1}^n w_i(x^i - x)^2}_{\text{bias}} \tag{7}$$

In the above, the expressions in red depend of f, those in blue depend on x and $x^{1:N}$.

Qualitative analysis of the bias terms

The first order term $f'(x)\sum_{i=1}^{n} w_i(x^i - x)$ is responsible for **border effects**.
The second order term **smooths out** sharp peaks and valleys.

Bias, Variance and h for $x \in \mathbb{R}$

[2]

The **bias** of $\hat{y}$ at x is defined as $E_{P_X^n} E_{P_\varepsilon^n}[\hat{y}(x) - f(x)]$.

$$E_{P_X^n} E_{P_\varepsilon^n}[\hat{y}(x) - f(x)] = h^2 \sigma_b^2 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right) + o(h^2) \tag{8}$$

The **variance** $\hat{y}$ at x is defined as $Var_{P_X^n P_\varepsilon^n}(\hat{y}(x))$.

$$Var_{P_X^n} P_\varepsilon^n(\hat{y}(x)) = \frac{\gamma^2}{nh}\sigma^2 + o\left(\frac{1}{nh}\right). \tag{9}$$

The **MSE (Mean Squared Error)** is defined as $E_{P_X^n} E_{P_\varepsilon^n}\left[(\hat{y}(x) - f(x))^2\right]$, which equals

$$MSE(x) = \textbf{bias}^2 + \textbf{variance}^2 = h^4 \sigma_b^4 \left(\frac{f'(x) p_X'(x)}{p_X(x)} + \frac{f''(x)}{2} \right) + \frac{\gamma_b^2}{nh}\sigma^2 + \ldots \tag{10}$$

[2] After []

Optimal selection of h

If the MSE is integrated over $\mathbb{R}$ we obtain the **MISE**$= \int_{\mathbb{R}} MSE(x) p_X(x) dx$.
The kernel width h can be chosen to minimize the MISE, for fixed f, p_X and b.
We set to 0 the partial derivative

$$\frac{\partial MISE}{\partial h} = h^3 \left(\text{[redacted]} \right) - \frac{(\text{[redacted]})}{nh^2} = 0. \tag{11}$$

It follows that $h^5 \propto \frac{1}{n}$, or

$$h \propto \frac{1}{n^{1/5}}. \tag{12}$$

In d dimensions, the optimal h depends on the sample size n as

$$h \propto \frac{1}{n^{1/(n+4)}}. \tag{13}$$