인공지능을 위한 이론과 모델링

2강: Nonparametric Regression

Lecturer: 장원철 *2023* 가을학기

2.1 Introduction

2.1.1 Definition

서로 독립이고 동일한 분포를 따르는 $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)$를 관측했다고 가정하자. 여기서 y는 scalar이고 $\mathbf{X} = (X_1, X_2, \ldots, X_p)$는 d차원 벡터이다. 우리는 다음과 같은 조건부 기댓값, 즉 회귀함수를 추정하는 것에 관심이 있다.

$$m_0 = \mathsf{E}\,(Y|\mathbf{X} = \mathbf{x})$$

여기서 ϵ_i와 X_i는 서로 독립이라고 가정하자. 사실 이 가정은 생각보다 굉장히 강한 가정이다. 예를 들면 $d = 10$인 경우, 즉 예측변수(predictor = feature)의 갯수가 10인 경우를 생각해보자. 예측변수간은 사실 일반적으로 독립이 아니다. (만약 독립이라면 각 변수별로 단순회귀분석모형을 적합한 후 그 결과를 다 합치면 된다!) 이 경우 우리가 적합한 모형에는 7개의 예측변수만 있다고 가정하자. 이렇게 될 경우 오차항에 결과 나머지 3개의 예측변수는 잔차 $(= y_i - \widehat{y}_i)$에 포함될 수 있으므로 결국 이 모형에서는 오차항과 예측변수가 서로 독립이 아니다!

이러한 가정을 하지 않는 경우로는 $\mathbf{X}$가 고정된 값을 가진다고 간주하는 경우이다. 예를 들면 Wavelets과 같은 경우 X_j가 가지는 값은 주어진 구간안에서 2^k갯수만큼 등간격으로 떨어져 있는 값들이다. 실제 앞으로 나올 내용은 X_j들이 확률변수인지 여부에 관계없이 다 성립하는 내용들이다.

2.1.2 Norm and Risk

모든 모형에서 얼마나 실제 데이터에 잘 적합하는지 판단하기 위해서는 일단 회귀함수와 회귀함수 추정치간의 거리를 정의해야 한다. 일반적으로 많이 사용되는 norm은 다음과 같다.

- Empirical norm $\|\cdot\|_n$:
$$\|m\|_n^2 = \frac{1}{n}\sum m^2(X_i)$$
- L_2 norm $\|\cdot\|_2$: assuming random inputs are from P_X, ,
$$\|m\|_2^2 = \mathsf{E}\,(m^2(X)) = \int m^2(x) dP_X(x)$$

그렇다면 추정량의 적합도 정도는 위의 2가지 norm을 사용하여 측정할 수 있다.

$$\|\widehat{m} - m_0\|_n^2 \text{ or } \|\widehat{m} - m_0\|_2^2$$

위와 같은 norm을 *(empirical) L_2 error*라고 부른다. 여기서 $\widehat{m}$은 데이터에 따라서 달라지는 값이기 때문에 확률변수로 생각할 수 있다. 따라서 L_2 error 역시 random이다. 만약 우리가 새로운 관측치 (X, Y)를 발견한다면 이 관측치의 L_2 error의 기댓값은 다음과 같다.

$$\mathsf{E}\,(Y - \widehat{m}(X))^2 = \int b_n^2 dP_X(x) + \int v(x) dP_X(x) + \tau^2 = \|\widehat{m} - m_0\|^2 + \tau^2$$

여기서 $b_n(x) = \mathsf{E}\,[\widehat{m}(x)] - m_0(x)$은 bias, $v(x) = \mathsf{Var}\,(\widehat{m}(x))$은 variance이며 $\tau^2 = \mathsf{E}\,(Y - m_0(x))^2$은 오차항 ϵ의 분산이다. 우리는 모형의 복잡도(complexity)를 조절하는 최적의 조절모수 (smoothing parameter) 찾고자 한다. 모형의 복잡도는 bias와 variance의 tradeoff로 이루어지며 그림1은 이 관계를 보여주고 있다.

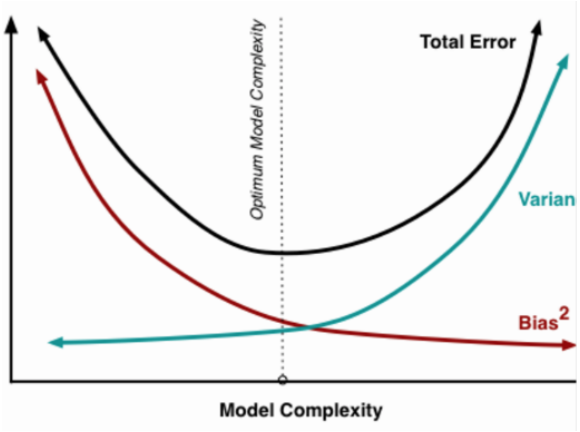

그림 1: Bias-variance tradeoff as a function of model complexity

2.1.3 비모수 추론 (nonparametric inference)이란?

비모수 추론이 의미하는 것은 크게 classical nonparametric inference과 modern inference 두가지로 나누어서 생각할 수 있다. 전자의 경우는 rank를 기반으로 하는 통계모형으로 Freedman's test, Kruskal-Wallis test, Wilcoxon rank test 등을 들 수 있다. 이 수업에서는 후자에 해당하는 modern nonparametric inference을 중점적으로 다룰 예정이다. 이 경우 nonparametric은 사실 모수의 갯수가 무한개임을 의미한다. Modern nonparametric inference은 함수추정에 있어서 괄목한 만한 성과물을 쏟아내고 있다. 비모수 추론에서 추정하고자 하는 함수는 특정 함수공간에 속한다는 가정을 한다. 보통 이런 함수공간은 거기에 속한 함수들이 얼마나 매끈한지를 보여주는 제약조건이라고 생각할 수 있다.

만약 $m_0(x)$가 Sobelov 공간에 속한다고 가정할 경우 적절한 basis function $g_j(x)$를 이용하여 다음과 같이 나타낼 수있다.

$$m_0 = \sum_{j=1}^{\infty} \beta_j g_j(x)$$

즉 m_0를 추정하는 것은 무한개의 모수 $\{\beta_j\}_{j=1}^{\infty}$를 추정하는 것과 같다. 하지만 우리가 데이터의 갯수보다 많은 모수를 추정할 수 없으므로 m_0의 근사함수 $\bar{m}(x) = \sum_{j=1}^{n} \beta_j g_j(x)$를 추정하는 것을 고려할 수 있다. 이 경우 추정량은 다음과 같다.

$$\widehat{m}(x) = \sum_{j=1}^{n} \widehat{\beta}_j g_j(x)$$

위의 추정량의 L_2 error는 다음과 같이 표현할 수 있다.

$$\|m_0 - \widehat{m}\|_2^2 = \underbrace{\|m_0 - \bar{m}\|_2^2}_{\text{approximation error}} + \underbrace{\|\bar{m} - m_0\|_2^2}_{\text{estimation error}}$$

여기서 첫번째 항은 근사오차(*approximation error*)를 나타내며 이 오차는 두번째항인 추정오차(*estimation error*)보다 훨씬 작다. 이 근사오차의 한계는 함수공간에 따라 정해진다. 다음 정리는 spline basis를 사용할 경우의 근사오차 최대치를 알려준다.

Theorem 2.1. (de Boor 1978) For any twice differentiable function m_0 on $[0,1]$ and points $t_1, \ldots, t_N \in [0,1]$, there is a cubic spline $\bar{m}_0$ with knots at $t_1, \ldots, t_N$ such that

$$\sup_{x \in [0,1]} |\bar{m}(x) - m_0(x)| \leq \frac{C}{N} \sqrt{\int m_0''(x)^2 dx}$$

With $N = \sqrt{n}$, one can easily see the approximation error is of order $1/n$ which is smaller than the estimation error in nonparametric function estimation.

2.1.4 Function Spaces

그렇다면 대표적인 함수공간은 어떤 것이 있을까? 이 절에서는 함수추정에서 많이 사용하는 함수공간들에 대해서 알아보자.

- The class of Lipschitz functions $H(1, L)$ on $T \subset \mathbb{R}$ is the set of functions g such that

$$|g(y) - g(x)| \leq L|x - y| \text{ for all } x, y \in T.$$

- The Hölder space $H(\beta, L)$ is the set of functions g mapping T to $\mathbb{R}$ such that g is $\ell = \beta - 1$ times differentiable and satisfies

$$|g^{\ell}(y) - g^{\ell}(y)| \leq L|x - y| \text{ for all } x, y \in T.$$

- The Sobolev space $S_1(\beta, L)$ is the set of β times differentiable functions $g : \mathbb{R} \to \mathbb{R}$ such that

$$\int (g^{(\beta)}(x))^2 dx \leq L^2.$$

2.2 Regressogram and k-nearest-neighbors regression

2.2.1 Regressogram

히스토그램은 아마 가장 보편적으로 사용되는 분포의 형태를 시각적으로 표현하는 방법이다. 히스토그램을 상대도수로 표현한다면 확률밀도함수(probability density function)을 추정할 수 있는 것과 같은 개념으로 회귀함수를 추정하는 방법이 regressogram이다.

먼저 우리가 $(X_1, Y_1), \ldots, (X_n, Y_n)$를 관측한다고 가정하자. 여기서 $X_i \in [0,1]$이고 Y는 bounded 되었다고 가정하자. 즉 $|Y| \leq C < \infty$ for some C. 추가로 X_i의 확률밀도함수 f가 다음과 같은 조건을 만족한다고 가정하자.

$$\inf_{x \in [0,1]} f(x) \geq c > 0$$

이제 m_0가 [0,1]에서 정의된 Lipschitz continuous 함수라고 하자. 이 경우 regressogram은 다음과 같이 정의된다.

$$\widehat{m}(x) = \frac{\sum_{i=1}^n Y_i I(X_i \in B_\ell)}{\sum_{i=1}^n I(X_i \in B_\ell)}, \text{ for } \ell = 1, \ldots, M$$

여기서 $B_\ell = [\frac{\ell-1}{M}, \frac{\ell}{M})$이며 M은 bin의 갯수를 나타낸다. 일반적으로 binwidth $h = 1/M$를 이용해서 위의 식을 다시 나타낼 수 있다. 여기서 M 또는 h가 smoothing parameter로 이 smoothing parameter를 optimal하고 정하는 것이 regressogram의 성능을 좌우하게 된다.

Optimal smoothing parametric 정하기 위해 일반적으로 사용하는 방법은 MSE(Mean Squared Error)를 계산한 후 이를 최소로 하는 smoothing parameter M^*를 찾는 것이다. 이 경우

$$\begin{aligned} \mathsf{E}\left((\widehat{m}(x) - m_0(x))^2\right) &= \underbrace{\left(\mathsf{E}\left[\widehat{m}(x)\right] - m_0(x)\right)^2}_{\text{Bias}^2(\widehat{m}(x))} + \underbrace{\mathsf{E}\left[(\widehat{m}(x) - \mathsf{E}\left[\widehat{m}(x)\right])\right]^2}_{\mathsf{Var}\left(\widehat{m}(x)\right))} \\ &= O\left(\frac{1}{M^2}\right) + O\left(\frac{M}{n}\right). \end{aligned}$$

따라서 $M^* \asymp n^{1/3}$이 되며M^*를 위의 MSE 공식에 대입하면 optimal convergence rate를 구할 수 있다. 이 경우 optimal convergence rate은 $O(n^{-2/3})$이다.

2.2.2 k-nearest-neighbors regression

만약 특정 bin안에 있는 데이터가 하나도 있지 않을 경우 해당 bin에서 regressogram의 추정치는 0이 된다. 이런 문제를 해결하기 위한 가장 쉬운방법은 bin의 갯수나 폭을 smoothing parameter로 하지 않고 bin안의 데이터의 개수를 smoothing parameter로 정하는 것이다. 즉 모든 bin에 들어가는 데이터의 개수는 동일하게 유지하기 위해 bin의 폭을 변화시키는 것이다.

k-nearest-neighbors (knn) regression은 다음과 같이 정의할 수 있다.

$$\widehat{m}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} Y_i \tag{1}$$

여기서 $\mathcal{N}_k(x)$은 $X_1, \ldots, X_n$중 x에 가장 가까운 k 자료의 index set을 나타낸다.

위의 식 (**??**)은 다음과 같이 표현할 수도 있다.

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i \tag{2}$$

여기서

$$w_i(x) = \begin{cases} \frac{1}{k} & \text{if } x_i \text{ is one of the } k \text{ nearest points to } x \\ 0 & \text{otherwise.} \end{cases}$$

knn에서 $k = k_n$이 smoothing parametric임을 알 수 있다.

식 (**??**)은 사실 knn이 Y_i의 선형결합임을 보여준다. 이런식으로 표현되는 회귀함수추정치를 *linear smoother*라고 한다. 사실 많은 nonparametric regression function 추정치가 linear smoother임을 보일 수 있고 이를 통하여 일반화된 다양한 이론적 성질을 증명할 수 있다. 우선 다음 두가지 가정 (i) $\mathsf{E}(Y^2) < \infty$;(2) $k_n \to \infty, k_n/n \to 0$ 하에서 knn 추정치는 *universally consistent* 이다.

$$\mathsf{E} \|\widehat{m} - m_0\|_2^2 \to 0, \text{ as } n \to \infty$$

뿐만 아니라 m_0 가 Lipschitz continuous일 경우 $k \asymp n^{2/(2+d)}$이라면 다음 결과를 보여줄 수 있다.

$$\mathsf{E} \|\widehat{m} - m_0\|_2^2 \lesssim n^{n^{-2/(2+d)}}$$

Proof.

$$\begin{aligned} \mathsf{E}(\widehat{m}(x) - m_0(x))^2 &= \underbrace{(\mathsf{E}[\widehat{m}(x)] - m_0(x))^2}_{\text{Bias}^2(\widehat{m}(x))} + \underbrace{\mathsf{E}[(\widehat{m}(x) - \mathsf{E}[\widehat{m}(x)])]^2}_{\mathsf{Var}(\widehat{m}(x)))} \\ &= \left(\frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} (m_0(X_i) - m_0(x)) \right)^2 + \frac{\sigma^2}{k} \\ &\leq \left(\frac{L}{k} \sum_{i \in \mathcal{N}_k(x)} \|X_i - x\|_2^2 \right)^2 + \frac{\sigma^2}{k} \end{aligned}$$

X_i기 등간격으로 배열되었다고 가정한다면 마지막 줄의 첫번째 항은 다음과 같음을 보일 수 있다. For $C > 0$,

$$\|X_i - x\|_2 \leq C(k/n)^{1/d}$$

따라서 MSE는 다음과 같이 정의된다,

$$(CL)^2 \left(\frac{k}{n} \right)^{2/d} + \frac{\sigma^2}{k}$$

여기서 bias-variance tradeoff를 사용한다면 optimal smoothing parameter는 $k \asymp n^{2/(2+d)}$이 된다 따라서 이 smoothing parameter를 MSE의 계산식에 다시 대입하면 optimal convergence rate가 $n^{-2/(2+d)}$임을 보일 수 있다. □

위의 optimal convergence rate은 차원의 저주 (curse of dimensionality)를 보여주고 있다. 데이터의 크기 n이 증가한다면 우리는 risk가 작아질 것을 기대한다. 선형회귀모형에서 risk의 convergence rate은 d/n이다. Risk의 크기를 ϵd이라고 한다면 knn에서 risk가 ϵ만큼 작아지기 위한 표본크기 n은 $\epsilon^{-(2+d)/d}$보다 크야 한다는 것을 알 수 있다. 그림 **??**은 $\epsilon = 0.1$일때 차원별로 필요한 표본크기를 보여주고 있다. 차원이 증가함에 따라 표본크기가 지수함수 형태로 증가함을 알 수 있다.

이러한 현상을 차원의 저주라고 하며 일반적으로 다음과 같은 부등식을 보여줄 수 있다.

$$\inf_{\widehat{m}} \sup_{m_0 \in H_d(1,L)} \mathsf{E} \|\widehat{m} - m_0\|_2^2 \gtrsim n^{-2/(2+d)},$$

여기서 $H_d(1, L)$는 d차원의 L-Lipschitz 함수공간을 나타낸다. 즉 이 함수공간에서는 knn보다 더 좋은 convergence rate을 가지는 추정방법은 없다. 그렇다면 우리가 보다 복잡한 함수도 고려할 수 있는 함수공간을 고려한다면 어떤 추정방법을 사용해야 할까? 그 질문에 대답하기 위해 kernel smoothing에 대해 알아보자.

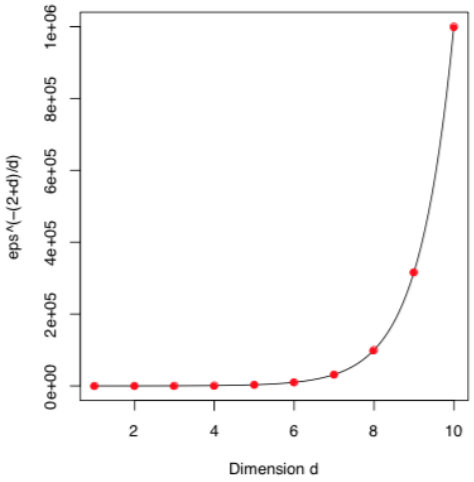

그림 2: The curse of dimensionality with $\epsilon = 0.1$

2.3 Kernel Smoothing and Local Polynomials

2.3.1 Kernel Smoothing

knn은 사실 실제로 많이 사용되지 않는데 그 이유는 k가 작을 경우 추정치가 상당히 울퉁불퉁하게 나타난다. 그 이유는 가중치 $w_i(x)$을 불연속성에 기인한 것이다. 그렇다면 추정치를 보다 매끈하게 보이게 하기 위해서는 가중치를 연속함수를 사용하는 것을 고려할 수 있고 이런 가중치를 *kernel*이라고 한다.

일반적으로 kernel 함수 $K : \mathbb{R} \to \mathbb{R}$는 다음과 같은 조건을 만족한다.

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad 0 < \int t^2 K(t) < \infty$$

보편적으로 많이 사용되는 kernel 함수는 다음 3가지이다.

- Box-car kernel:
$$K(t) = \begin{cases} 1 & \text{if } |t| \leq 1/2. \\ 0 & \text{otherwise} \end{cases}$$
- Gaussian kernel:
$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$
- Epanechnikov kernel:
$$K(t) = \begin{cases} 3/4(1-t^2) & |t| \leq 1. \\ 0 & \text{otherwise} \end{cases}$$

knn을 일반화한 Nadayara-Watson kernel regression은 다음과 같이 정의된다.

$$\widehat{m}(x) = \frac{\sum_{i=1}^n K\left(\frac{\|x-X_i\|}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{\|x-X_i\|}{h}\right)} = \sum_{i=1}^n w_i(x) Y_i$$

여기서 $w_i(x) = K\left(\frac{\|x-X_i\|}{h}\right) / \sum_{i=1}^n K\left(\frac{\|x-X_i\|}{h}\right)$이다. 그림 **??**는 1차원에서 knn과 Nadayara-Watson regression의 추정치를 보여주고 있다.

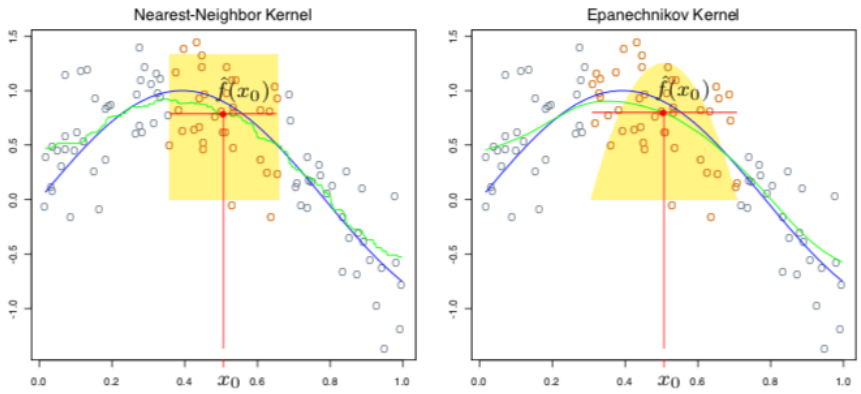

그림 3: knn vs kernel regression with Epanechnikov kernel when $d = 1$ (Hastie et al. 2009)

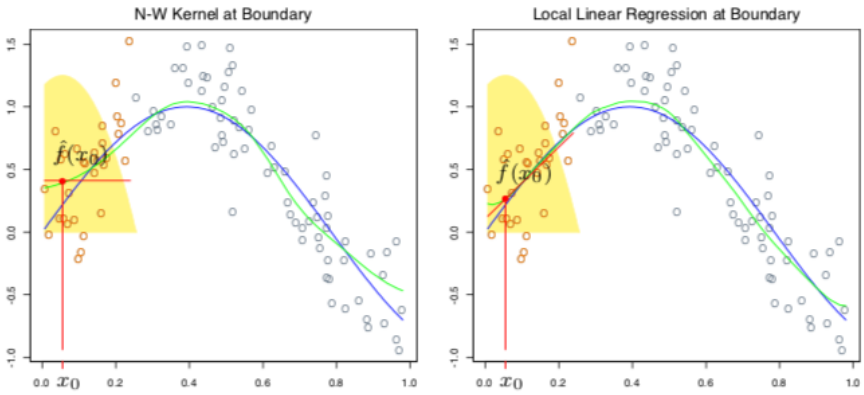

그림 4: kernel regression vs local regression with Epanechnikov kernel when $d = 1$ (Hastie et al. 2009)

만약 compactly supported kernel K과 bandwidth $h = h_n$가 n이 무한대로 갈 경우 $h_n \to 0$ and $nh_n^d \to \infty$ 이라면 kernel regression estimator 역시 *universally consistent*하다.

$$\mathsf{E}\,\|\widehat{m} - m_0\|_2^2 \to 0 \text{ as } n \to 0,$$

보다 자세한 내용은 다음 정리를 참조하자.

Theorem 2.2. Suppose that $d = 1$ and that m'' is bounded. Also suppose that X has a non-zero, differentiable density p and the support is unbounded. Then, the risk is

$$\begin{aligned} R_n &= \frac{h_n^4}{4}\left(\int x^2 K(x)dx\right)^2 \int \left(m''(x) + 2m'(x)\frac{p'(x)}{p(x)}\right)^2 dx \\ &+ \frac{\sigma^2 \int K^2(x)dx}{nh_n}\int \frac{1}{p(x)}dx + o\left(\frac{1}{nh_n}\right) + o(h_n^4). \end{aligned}$$

위의 식에서 첫번째항은 bias^{2}을 나타내고 두번째항은 분산항에서 나온 것이다. 이 경우 optimal bandwidth의 order는 $O(n^{-1/5})$이며 여기에 대응되는 risk (optimal convergence rate)의 order는 $O(n^{-4/5})$. 여기서 또 하나 주목할 점은 boundary에서 bias의 order가 $O(h)$로 증가한다는 점이다. 그림 **??**은 kernel smoothing을 사용할 경우 boundary에서 문제점을 보여준다. Local polynomials을 사용하면 이 문제를 해결할 수 있다.

2.3.2 Local Polynomials

p차 다항식을 사용하는 회귀모형을 생각해보자.

$$\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 x^2 + \cdots + \widehat{\beta}_p x^p$$

여기서 $\widehat{\beta} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$이며 최소제곱법을 통해서 구할 수 있다.

Local polynomial regression은 모든 x에 국소적으로 다른 다항회귀를 적합시켜 $f(x)$ 를 근사하는 방법이다.

$$f(u) \approx \beta_0(x) + \beta_1(x)(u-x) + \beta_2(u-x)^2 + \cdots \beta_p(x)(u-x)^p$$

여기서 u는 x의 인근의 점이다.

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - [\widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_2 x^2 + \cdots + \widehat{\beta}_p x^p] \right)^2$$

위의 식에서 $\widehat{f}(x) = \widehat{\beta}_0(x)$!를 이용하여 구할 수 있다. 또한 $p = 0$경우는 kernel regression이 됨을 알수 있다.

$$\begin{aligned} \widehat{f}(x) &= \sum_{i=1}^{n} \ell_i(x) Y_i \\ \ell_i(x) &= \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)} \end{aligned}$$

만액 $p = 1$이라면 **local linear estimator**가 되고 대부분의 경우 local linear estimator가 아주 훌륭한 추정량으로 사용할 수 있다. 모든 kernel method에서 기억해야 할 사항은 kernel의 종류는 중요하지 않고 smoothing parameter h를 어떻게 선택하는냐가 추정치가 얼마나 정확한지를 결정한다는 사실이다.

Local polynomial regression estimate는 다음과 같이 주어진다.

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

여기서 $\ell(x)^T = (\ell_1(x), \ldots, \ell_n(x))$, 즉

$$\ell(x)^T = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x,$$

$e_1 = (1, 0, \ldots, 0)^T$이며 X_x와 W_x는 다음과 같이 정의된다.

$$X_x = \begin{pmatrix} 1 & X_1 - x & \cdots & (X_1-x)^p/p! \\ 1 & X_2 - x & \cdots & (X_2-x)^p/p! \\ \vdots & \vdots & \vdots & \\ 1 & X_n - x & \cdots & (X_n-x)^p/p! \end{pmatrix}, W_x = \begin{pmatrix} K\left(\frac{x-X_1}{h}\right) & 0 & \cdots & 0 \\ 0 & K\left(\frac{x-X_2}{h}\right) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & K\left(\frac{x-X_n}{h}\right) \end{pmatrix}$$

Theorem 2.3. When $p = 1$, $\widehat{f} = \sum_{i=1}^{n} \ell_i(x) Y_i$, where

$$\ell_i(x) = \frac{b_i(x)}{\sum_{j=1}^{n} b_j(x)}.$$

Here

$$\begin{aligned} b_i(x) &= K\left(\frac{X_i - x}{h}\right)(S_{n,2}(x) - (X_i - x)S_{n,1}(x)), \\ S_{n,j}(x) &= \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x)^j, \quad j = 1, 2. \end{aligned}$$

2.4 Linear Smoothers

2.4.1 Definition

이때까지 소개한 모든 추정량은 linear smoother이다. 즉 Y_i들의 선형결합으로 표현할 수 있는 추정량이라는 의미이다.

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i.$$

여기서 $w_i(x)$는 x값에 의해 결정되고 모든 추정량은 smoothing parameter h가 있다.

$\widehat{\mu} = (\widehat{m}(X_1), \ldots, \widehat{m}(X_n))$라고 한다면 $\widehat{\mu} = Sy$으로 표현할 수 있으며 여기서 S는 smoothing matrix이다. 또한 *effective degrees of freedom*을 다음과 같이 정의할 수 있다.

$$\nu = \mathrm{df}(\widehat{\mu}) = \sum_{i=1}^{n} S_{ii} = tr(S)$$

Effective degrees of freedom은 사실 smoothing parameter h의 함수로 표현할 수 있으며 모형의 복잡도 (complexity)를 나타내는 척도로 사용할 수 있다. Effective degrees of freedom은 모든 linear smoother에서 정의할 수 있으므로 spline smoothing과 local regression처럼 서로 다른 smoothing parameter를 사용하더라도 모형의 복잡도를 비교할 수 있다는 장점을 가지고 있다.

2.4.2 Choosing the Smoothing Parameter

다음과 같은 training error의 기대값을 고려해보자.

$$\begin{aligned} \mathsf{E}\,(\text{traning error}) &= \mathsf{E}\left[\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}\,(Y_i - m(X_i) + m(X_i) - \widehat{m}(X_i))^2 \\ &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}(\widehat{m}(X_i)) - \frac{1}{n}\sum_{i=1}^{n}\underbrace{\mathsf{Cov}\,(Y_i, \widehat{m}(X_i))}_{=S_{ii}\sigma^2} \end{aligned}$$

일반적으로 training data에서 Y_i와 $\widehat{Y}_i = \widehat{m}(X_i)$는 양의 상관관계를 가지고 있다. 그래서 위의 식에서 3번째항이 training data가 risk를 과소추정하는 이유를 설명하고 있다. 3번째 항은 다음과 같이 표현할 수 있다.

$$\sum_{i=1}^{n} \mathsf{Cov}\left(Y_i, \widehat{m}(X_i)\right) = \mathrm{tr}(\mathsf{Cov}\left(Y, \widehat{\mu}\right)) = \mathrm{tr}(S\mathsf{Var}\left(Y\right)) = \sigma^2 \mathrm{tr}(S)$$

과소추정을 피하기 위해 다음과 같은 leave-one-out cross-validation을 고려해보자.

$$\mathrm{CV}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}^{-i}(X_i))^2$$

여기서 $\widehat{m}^{-i}$는 전체자료에서 ith data pair (X_i, Y_i)를 제외하고 계산한 추정량을 의미한다. 따라서 Y_i와 $\widehat{m}^{-i}(X_i)$는 서로 독립이기 때문에 3번째항에 대한 걱정은 하지 않아도 된다.

are independent so we don't need to worry about the third term.

또한 아래의 식을 사용한다면 손쉽게 leave-one-out CV을 계산할 수 있다.

$$\mathrm{CV}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}(X_i)}{1 - S_{ii}}\right)^2$$

Proof. Recall $\widehat{m}(X_i) = \sum_{i=1}^{n} S_{ij}Y_j$ where S_{ij} are (i, j) element of S. To remove the influence of the ith pair, consider

$$\widehat{m}(X_i) - S_{ii}Y_i = \sum_{i \neq j} S_{ij}Y_j$$

By renormalizing LHS,

$$\widehat{m}^{-i}(X_i) = \frac{1}{1 - S_{ii}}(\widehat{m}(X_i) - S_{ii}Y_i).$$

Hence

$$Y_i - \widehat{m}^{-i}(X_i) = Y_i - \frac{1}{1 - S_{ii}}(\widehat{m}(X_i) - S_{ii}Y_i) = \frac{Y_i - \widehat{m}(X_i)}{1 - S_{ii}}.$$

□

Generalized cross-validation (GCV) 는 leave-one-out cross-validation을 간편화 한 방식이다. 위의 식에서 S_{ii}를 $\sum_{i=1}^{n} S_{ii}/n = \mathrm{tr}(S)/n$로 대치하면 다음과 같은 식을 얻을 수 있다.

$$\mathrm{GCV}(\widehat{m}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{m}(X_i)}{1 - \nu/n}\right)^2.$$

여기서 ν는 effective degrees of freedom이다.

x가 작을 경우 $(1 - x)^{-2} \approx 1 + 2x$라는 사실을 이용한다면,

$$GCV(\widehat{m}) = \left(1 - \frac{\nu}{n}\right)^{-2} CV(\widehat{m}) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2 + \frac{2\nu}{n}\widehat{\sigma}^2$$

여기서 $\widehat{\sigma}^2$은 MSE (training error)를 이용하여 추정한다.

즉 $\widehat{\sigma}^2$ 를 full model (모든 예측변수를 포함한 모형)에서 계산된 MSE로 대치하면 위의 근사치는 잘 알려진 *Mallow's* C_p와 같다.

$$C_p = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2 + \frac{2\nu}{n}\widehat{\sigma}^2$$

여기서 $\widehat{\sigma}^2$ 는 full 모형의 MSE이다.