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 p차원 벡터이고다. 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 $p = 10$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 회귀함수외 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 $b_n(x) = \mathsf{E}\,[\widehat{m}(x)] - m(x)$은 bias, $v(x) = \mathsf{Var}\,(\widehat{m}(x))$은 variance이며 $\tau^2 = \mathsf{E}\,(Y - m(x))^2$은 오차항 ϵ의 분산이다. 우리는 모형의 복잡도(complexity)를 조절하는 최적의 조절모수 (smoothing parameter) 찾는데 있다. 모형의 복잡도는 bias와 variance의 tradeoff로 이루어지며 그림1은 이 관계를 보여주고 있다.

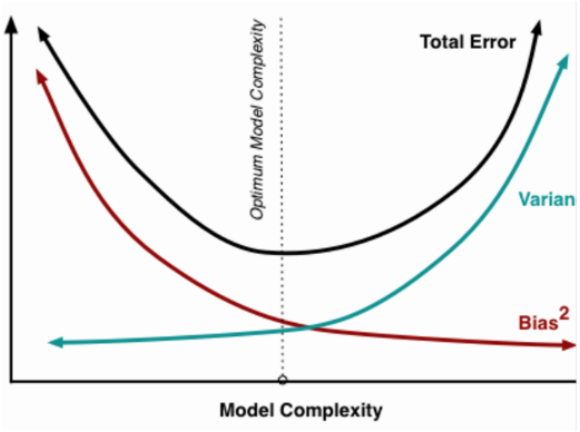

그림 1: Bias-variance tradeoff as a function of model complexity

2.1.3 Nonparametric 이란?

Nonparametric statistics이 의미하는 것은 크게 classical nonparametric statistics과 modern nonparametric statistics 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 statistics을 중점적으로 다룰 예정이다. 이 경우 nonparametric은 사실 모수의 갯수가 무한개임을 의미한다. Modern nonparametric statistics은 함수추정에 있어서 괄목한 만한 성과물을 쏟아내고 있다. 하지만 이러한 nonparametric inference에서 추정하고자 하는 함수가 특정 함수공간에 속한다는 가정을 한다. 보통 이런 함수공간의은 거기에 속한 함수들이 얼마나 smooth한지를 보여주는 제약조건이라고 생각할 수 있다.

만약 m_0가 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.

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

이 경우 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 많이 사용되는 분포의 형태를 시각적으로 표현하는 방법이다. 히스토그램을 상대도수를 이용하여 나탸낸다면 확률밀도함수(probability density function)을 추정치이다. 이와 유사한 개념으로 회귀함수의 추정치로 사용하는 것이 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, 종종 binwidth$h$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 parameter를 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 되면 the optimal convergence rate는 $O(n^{-2/3})$이다.

2.2.2 k-nearest-neighbors regression

Regressogram의 경우 특정 bin안에 있는 데이터의 개수가 굉장히 작을 수 있다. 따라서 이런 경우는 bin의 갯수나 폭을 smoothing parameter로 하지 않고 bin안의 데이터의 개수를 smoothing parameter로 정할 수 있다. 즉 각 bin의 폭을 변화시키면서 해당 bin에 들어가는 데이터의 개수는 동일하게 유지하는 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 (1)은 다음과 같이 표현할 수도 있다.

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

where

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

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

식 (2)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.

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

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.

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

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,

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

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. □