Lecture 3/4 Bias-Variance Tradeoff

IEMS 402 Statistical Learning

Northwestern

Announcement

max(HW1,HW8)+max(HW2,HW3)+max(HW4,HW5)+max(HW6,HW7).

Postpone for one week!

~~DDL 1.17~~
~~DDL 1.24~~

DDL 1.24

- [Homework 1] Review of Probability and Optimization
- [Homework 2] Bias and Variance Trade-off 1 — Easy
- [Homework 3] Bias and Variance Trade-off 2
- [Homework 4] Asymptotic Theory 1 — Easy
- [Homework 5] Asymptotic Theory 2
- [Homework 6] Non-Asymptotic Theory 1 — Easy
- [Homework 7] Non-Asymptotic Theory 2
- [Homework 8] Advanced Topics

- Latex and overleaf (not required)

Lecture note

https://www.stat.cmu.edu/~larry/=sml/nonpar2019.pdf
https://www.stat.cmu.edu/~larry/=sml/densityestimation.pdf

Local Smoothing

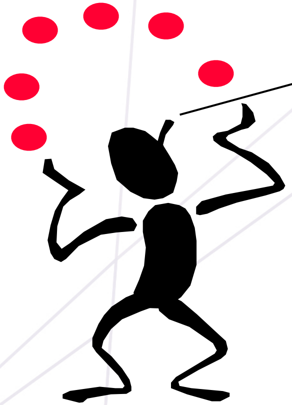
Non-parametric Regression

- The aim of a regression analysis is to produce a reasonable analysis to the unknown response function m, where for n data points $(x_i, y_i)_{i=1}^{n}$, the relationship can be modeled as

$$y_i = f(x_i) + \eta_i, \eta_i \sim N(0,1)$$

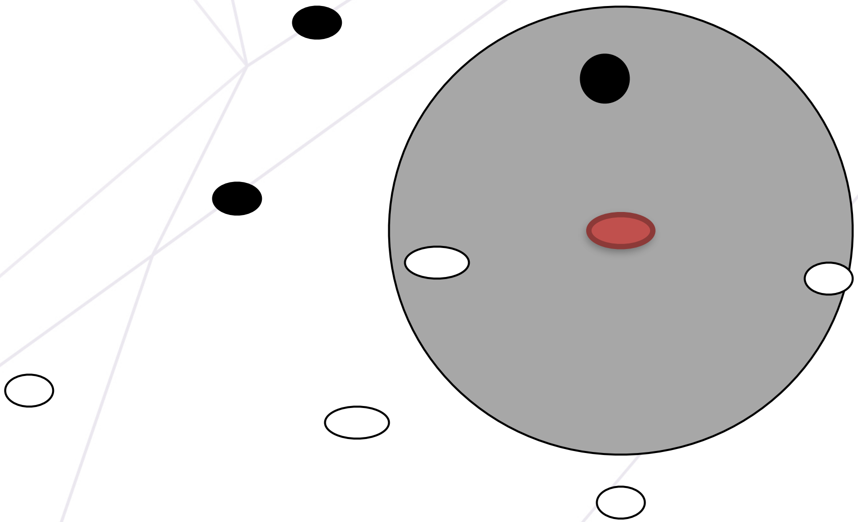
- Unlike parametric approach where the function m is fully described by a finite set of parameters, nonparametric modeling accommodate a very flexible form of the regression curve.

Instance-based learning



Its very similar to a Desktop!!

3-nearest neighbor



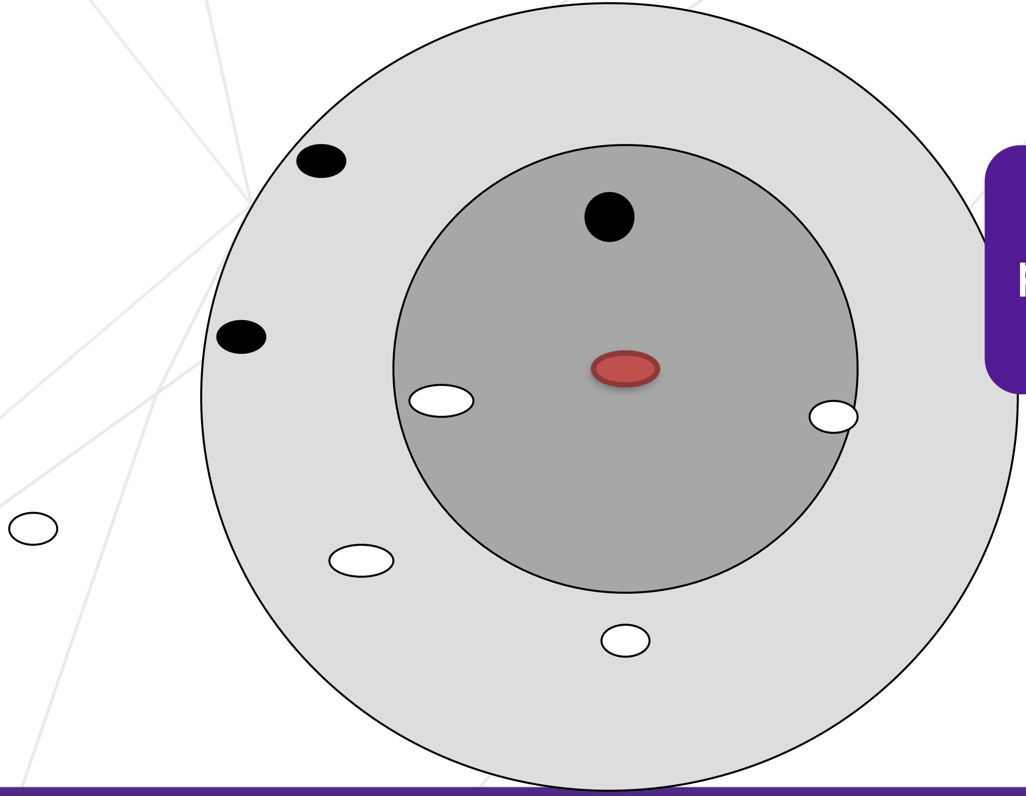
K-nearest Neighbor

Here's a basic method to start us off: *k-nearest-neighbors* regression. We fix an integer $k \geq 1$ and define

$$\widehat{m}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} Y_i, \tag{4}$$

where $\mathcal{N}_k(x)$ contains the indices of the k closest points of $X_1, \ldots, X_n$ to x.

Bias and Variance in k-NN



More data points but less similar data...

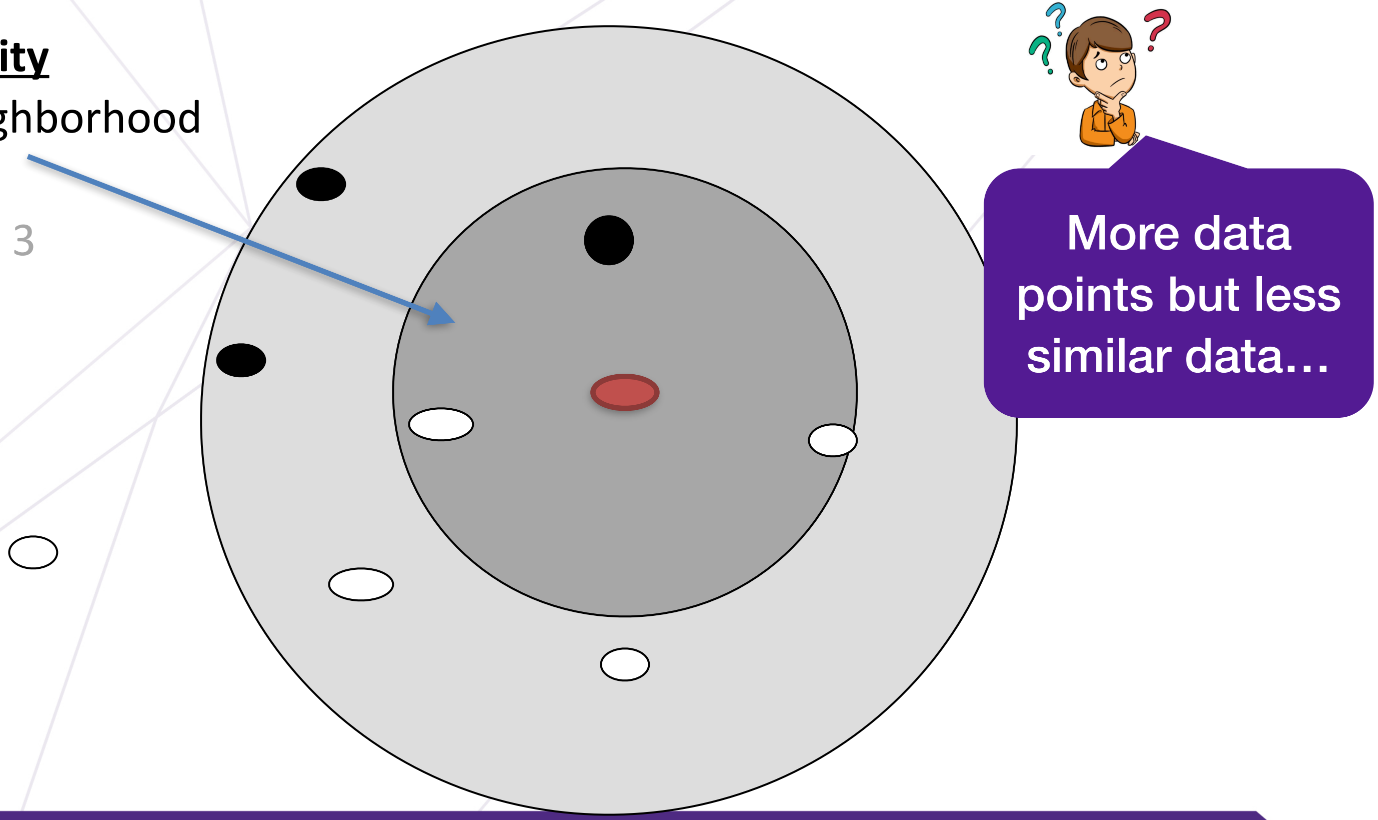
Bias and Variance in k-NN

Curse of Dimensionality

Fewer data in the neighborhood

In high dimension

Homework 1 Problem 3



Selecting k in k-NN

$$\mathbb{E}\left[\left(\widehat{m}(x) - m_0(x)\right)^2\right] = \underbrace{\left(\mathbb{E}[\widehat{m}(x)] - m_0(x)\right)^2}_{\mathrm{Bias}^2(\widehat{m}(x))} + \underbrace{\mathbb{E}\left[\left(\widehat{m}(x) - \mathbb{E}[\widehat{m}(x)]\right)^2\right]}_{\mathrm{Var}(\widehat{m}(x))}$$

$$= \left(\frac{1}{k}\sum_{i\in\mathcal{N}_k(x)} \left(m_0(X_i) - m_0(x)\right)\right)^2 + \frac{\sigma^2}{k}$$

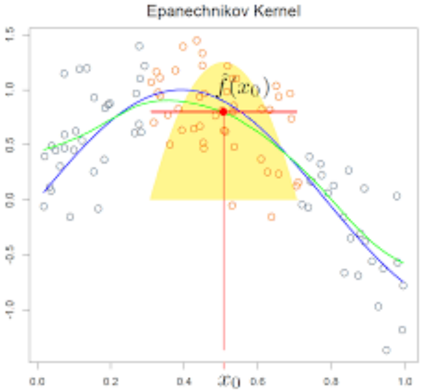
$$\leq \left(\boxed{\frac{L}{k}\sum_{i\in\mathcal{N}_k(x)} \|X_i - x\|_2}\right)^2 + \frac{\sigma^2}{k}.$$

$$\approx \left(\frac{k}{n}\right)^d$$

Homework 1 Problem 3

Local Averaging Procedure

- A reasonable approximation to the regression curve $m(x)$ will be the mean of response variables near a point x. This *local averaging procedure* can be defined as



$$\hat{m}(x) = n^{-1} \boxed{\sum_{i=1}^{n}} W_{ni}(x) Y_i \qquad (2)$$

Average out the noise!

Kernel Smoothing

The local averaging weights depend on the distance

$$W_{hi}(x) = K_h(x - X_i) / \boxed{\hat{f}_h(x)} \quad (3)$$

Normalize to be averaging!

Here $\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$

Kernel Smoothing

The local averaging weights depend on the distance

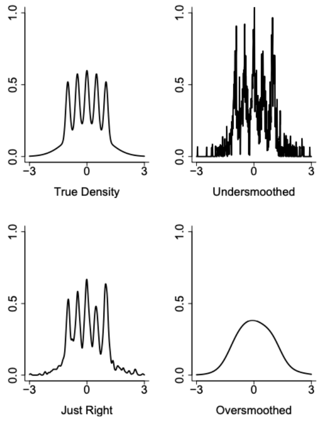
$$W_{hi}(x) = \boxed{K_h(x - X_i)} / \hat{f}_h(x) \qquad (3)$$

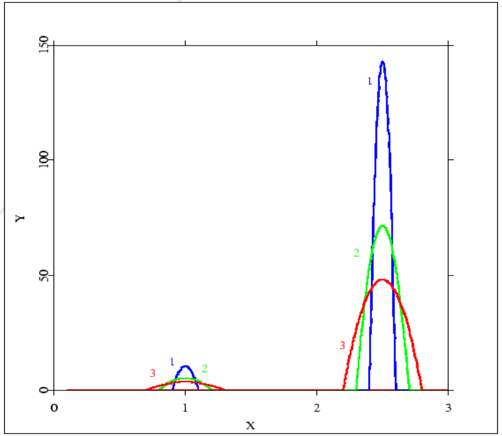
$$K_h(u) = h^{-d} K(u/h)$$

Here $\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$

h controls the size of the neighborhood!

Why -d ?





Kernel Smoothing

The local averaging weights depend on the distance

$$W_{hi}(x) = \boxed{K_h(x - X_i)} / \hat{f}_h(x) \qquad (3)$$

$$K_h(u) = h^{-d} K(u/h)$$

Here $\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$

h controls the size of the neighborhood!

- The *Nadaraya-Watson estimator* is defined by

$$\hat{m}_h(x) = \frac{n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i}{n^{-1} \sum_{i=1}^{n} K_h(x - X_i)} \qquad (4)$$

Error Analysis

Theorem: Risk bound without density. Suppose that the distribution of X has compact support and that $\mathrm{Var}(Y|X=x) \le \sigma^2 < \infty$ for all x. Then

$$\sup_{P \in H_d(1,L)} \mathbb{E}\|\widehat{m} - m\|_P^2 \le c_1 h^2 + \frac{c_2}{nh^d}. \tag{9}$$

Not Required

Hard, do a simpler model

Density Estimation

Kernel Density Estimation

Let $X_1, X_2, \cdots, X_n$ be a sample from a distribution P with density p. The goal of nonparametric density estimation is to estimate p with as few assumptions about p as possible.

Kernel Density Estimator:

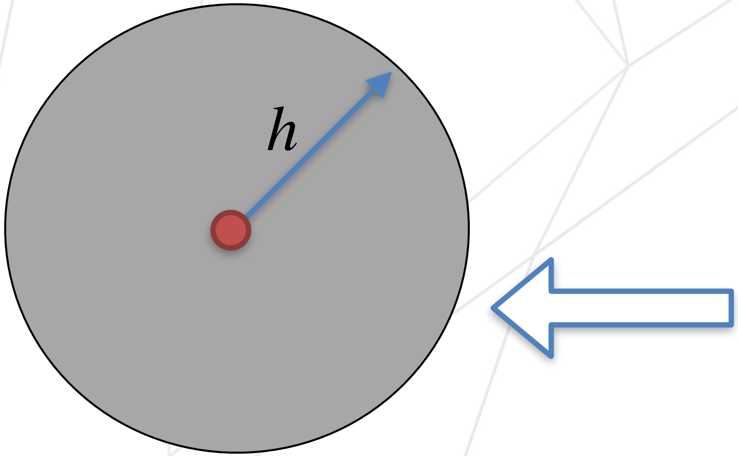
$$\widehat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K\left(\frac{\|x - X_i\|}{h}\right).$$

Homework 2 Problem 1 show an equivalence between Kernel Density Estimator and Kernel smoothing

Regards the bias

Consider an easier estimator $\widehat{p}_h(x) = \sum_{j=1}^{N} \frac{\widehat{\theta}_j}{h^d} I(x \in B_j)$

How histogram approximate the density



The volume is h^d

Regards the Variance

Consider an easier estimator $\widehat{p}_h(x) = \sum_{j=1}^{N} \frac{\widehat{\theta}_j}{h^d} I(x \in B_j)$

How histogram approximate the density

Recall

<u>Fact.</u> The number of parameters N required to achieve an approximation error of at most ϵ can be estimated by:

$$N \approx \left(\frac{1}{\epsilon}\right)^{\frac{\boxed{d}}{\boxed{s}}}$$

Dimension
smoothness

How can the smoothness helps?

What is the assumption behind…

Local regression: choices

Depend on the smoothness of target function

Choice 1: Type of model

- Linear regression
- Degree 2 polynomial
- Degree 3 polynomial

Choice 2: Weighting scheme

- Normal density
- Other schemes (called **kernels**)

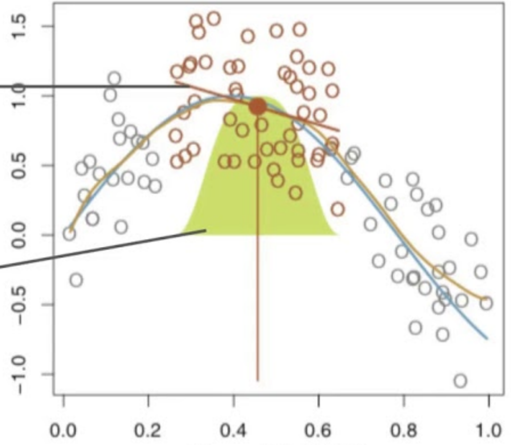


Figure 7.9 (ISLR)

What does linear mean

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)\boxed{Y_i}}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)}$$

The estimation is a linear function in Y

What does linear mean

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)\boxed{Y_i}}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)}$$

The estimation is a linear function in Y

Linear regression over a b c

How to do quadratic regression? $(X_i, Y_i)_{i=1}^n, Y_i \approx \boxed{a}X_i^2 + \boxed{b}X_i + \boxed{c}$

$$\boxed{\begin{bmatrix} a \\ b \\ c \end{bmatrix}} = \begin{bmatrix} 1 & X_1 & X_1^2 \\ \cdots & & \\ 1 & X_2 & X_2^2 \\ 1 & X_n & X_n^2 \end{bmatrix}^{\dagger} \boxed{\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}}$$

"Feature extraction"
Lecture 15

a, b, c is linear in y

All quadratic function forms a (linear) vector space!

What does linear mean

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)\boxed{Y_i}}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)}$$

The estimation is a linear function in Y

Linear regression over a b c

How to do quadratic regression? $(X_i, Y_i)_{i=1}^n, Y_i \approx \boxed{a}X_i^2 + \boxed{b}X_i + \boxed{c}$

$$\boxed{\begin{bmatrix} a \\ b \\ c \end{bmatrix}} = \begin{bmatrix} 1 & X_1 & X_1^2 \\ \cdots & & \\ 1 & X_2 & X_2^2 \\ 1 & X_n & X_n^2 \end{bmatrix}^{\dagger} \boxed{\begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}}$$

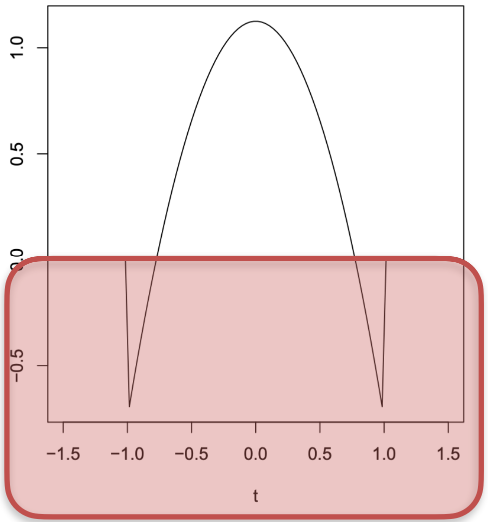
a, b, c is linear in y

Linear smoothing = local poly regression

All quadratic function forms a (linear) vector space!

Higher-order Kernel

$$\int K(t)\,dt = 1, \quad \int t^j K(t)\,dt = 0, \quad j = 1, \ldots, k-1, \quad \text{and} \quad 0 < \int t^k K(t)\,dt < \infty.$$



Local Regression vs Local Smoothing

Bias of local smoothing: $\int \underbrace{K_h(x - x_0)p(x)}[f(x) - f(x_0)]dx$

Need to cancel the Taylor expansion

We don't know what is the distribution p

Bias

Lemma 3 *The bias of $\widehat{p}_h$ satisfies:*

$$\sup_{p \in \Sigma(\beta, L)} |p_h(x) - p(x)| \leq ch^{\beta} \tag{14}$$

for some c.

Proof. We have

$$\begin{aligned}
|p_h(x) - p(x)| &= \int \frac{1}{h^d} K(\|u - x\|/h) p(u) du - p(x) \\
&= \left| \int K(\|v\|)(p(x + hv) - p(x)) dv \right| \\
&\leq \left| \int K(\|v\|)(p(x + hv) - p_{x,\beta}(x + hv)) dv \right| + \left| \int K(\|v\|)(p_{x,\beta}(x + hv) - p(x)) dv \right|.
\end{aligned}$$

The first term is bounded by $Lh^{\beta} \int K(s)|s|^{\beta}$ since $p \in \Sigma(\beta, L)$. The second term is 0 from the properties on K since $p_{x,\beta}(x + hv) - p(x)$ is a polynomial of degree β (with no constant term). $\square$

Variance

Lemma 4 *The variance of $\widehat{p}_h$ satisfies:*

$$\sup_{p\in\Sigma(\beta,L)} \mathrm{Var}(\widehat{p}_h(x)) \le \frac{c}{nh^d} \tag{15}$$

for some $c > 0$.

Proof. We can write $\widehat{p}(x) = n^{-1}\sum_{i=1}^n Z_i$ where $Z_i = \frac{1}{h^d} K\left(\frac{\|x-X_i\|}{h}\right)$. Then,

$$\begin{aligned}
\mathrm{Var}(Z_i) &\le \mathbb{E}(Z_i^2) = \frac{1}{h^{2d}} \int K^2\left(\frac{\|x-u\|}{h}\right) p(u)du = \frac{h^d}{h^{2d}} \int K^2\left(\|v\|\right) p(x+hv)dv \\
&\le \frac{\sup_x p(x)}{h^d} \int K^2(\|v\|)dv \le \frac{c}{h^d}
\end{aligned}$$

for some c since the densities in $\Sigma(\beta, L)$ are uniformly bounded. The result follows. □

Final Result

The optimal bound one can get

$$\sup_{p \in \Sigma(\beta,L)} \mathbb{E} \int (\widehat{p}_h(x) - p(x))^2 dx \preceq \left(\frac{1}{n}\right)^{\boxed{\frac{2\beta}{2\beta+d}}}.$$

Estimating the derivatives

Given a kernel function $K : \mathbb{R} \to \mathbb{R}$ supported on $[-1, 1]$ satisfying the conditions

$$\int_{\mathbb{R}} u^j K(u) du = \begin{cases} 1 & j = 1, \\ 0 & j = 0, 2, \cdots, \lfloor \beta \rfloor. \end{cases}$$

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p$. Given bandwidth $h > 0$, consider the kernel-based estimator

$$\widehat{d}_n(x) := \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

For any x_0, and prove the MSE bound

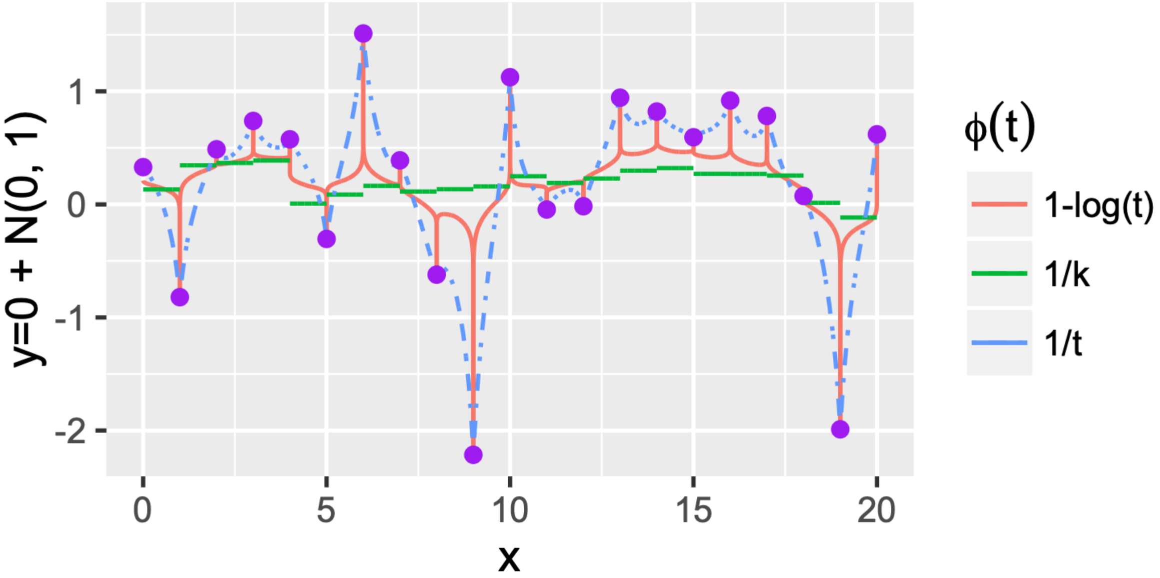
$$\mathbb{E}\left[|\widehat{d}_n(x_0) - p'(x_0)|^2\right] \leq n^{-\frac{\boxed{2(\beta-1)}}{1+2\beta}}.$$

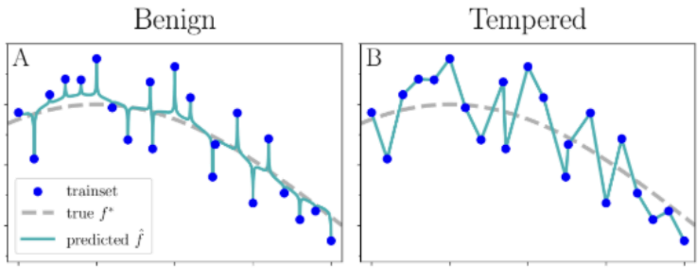
with an optimal bandwidth $h = h_n$

Not Required

Estimating the derivatives

Ok… Interpolation…(1-NN)





Xing Y, Song Q, Cheng G. Benefit of interpolation in nearest neighbor algorithms. SIAM Journal on Mathematics of Data Science, 2022, 4(2): 935-956.