

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)

Local Smoothing

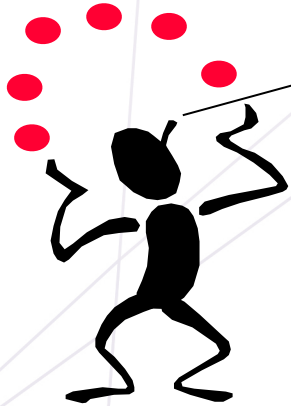
Non-parametric Regression

- 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.

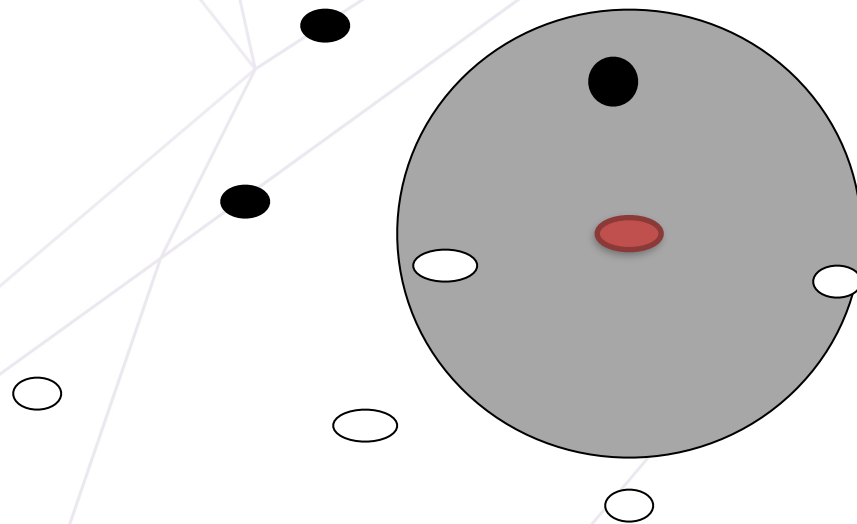
Instance-based learning



Its very similar to a
Desktop!!



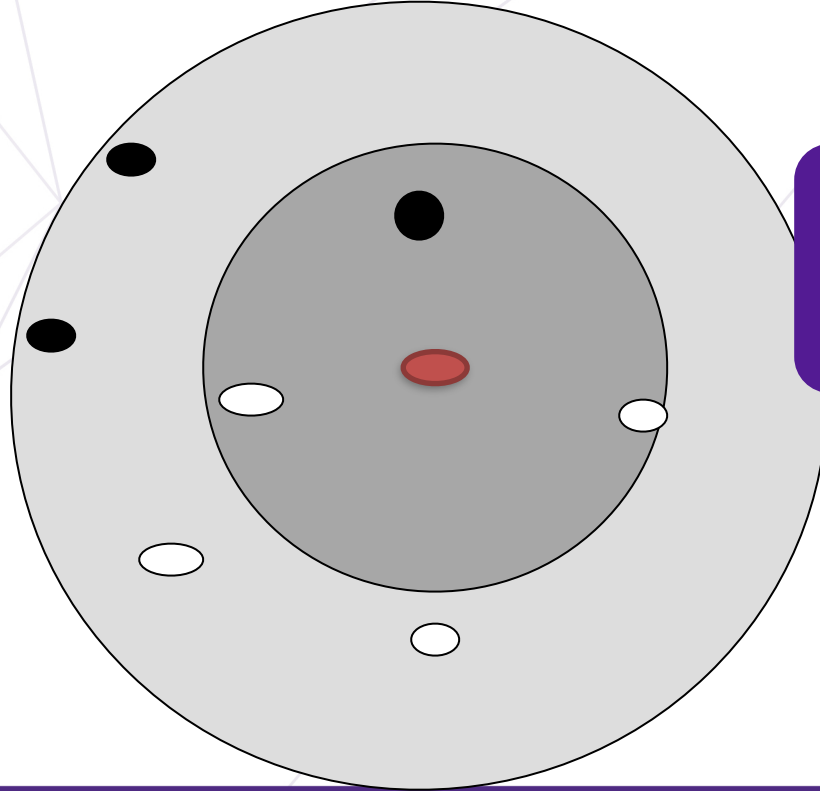
3-nearest neighbor



Bias and Variance in k-NN



More data
points but less
similar data...



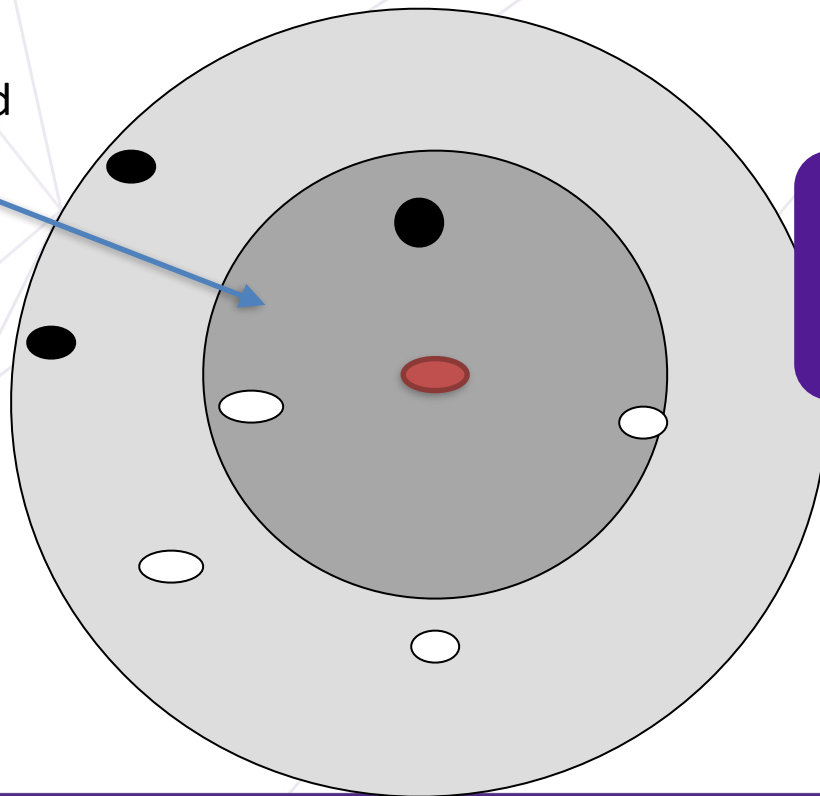
Bias and Variance in k-NN

Curse of Dimensionality

Fewer data in the neighborhood

In high dimension

Homework 1 Problem 3



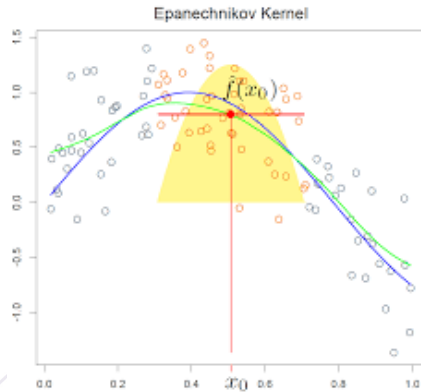
More data
points but less
similar data...

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

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

Average out the noise!



Kernel Smoothing

The local averaging weights depend on the distance

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

Normalize to be averaging!

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

Kernel Smoothing

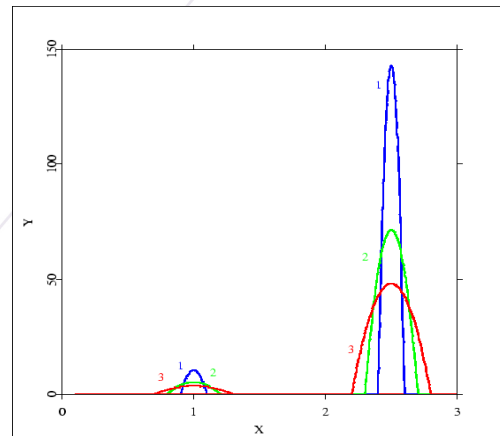
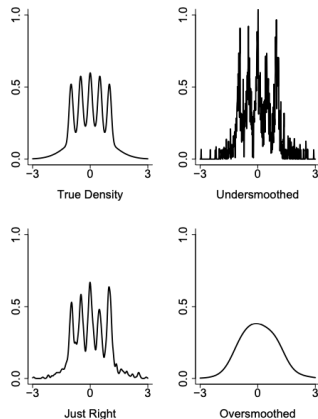
The local averaging weights depend on the distance

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

$$K_h(u) = h^{-d} K(u/h) \quad \text{Here } \hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$$

h controls the size of the neighborhood!

Why -d ?



Kernel Smoothing

The local averaging weights depend on the distance

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

$$K_h(u) = h^{-1} K(u/h) \quad \text{Here } \hat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$$

h controls the size of the neighborhood!

- The *Nadaraya-Watson estimator* is defined by

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

Selecting k in k-NN

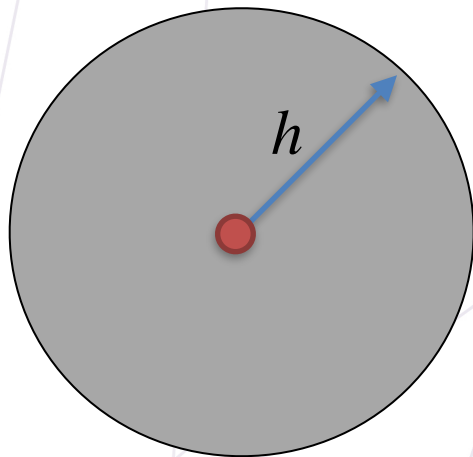
$$\begin{aligned}\mathbb{E}[(\hat{m}(x) - m_0(x))^2] &= \underbrace{(\mathbb{E}[\hat{m}(x)] - m_0(x))^2}_{\text{Bias}^2(\hat{m}(x))} + \underbrace{\mathbb{E}[(\hat{m}(x) - \mathbb{E}[\hat{m}(x)])^2]}_{\text{Var}(\hat{m}(x))} \\ &= \left(\frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} (m_0(X_i) - m_0(x)) \right)^2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

Homework 1 Problem 3

Regards the bias

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

How histogram approximate the density



The volume is h^d

Regards the Variance



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

How histogram approximate the density

Recall

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

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

 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**)

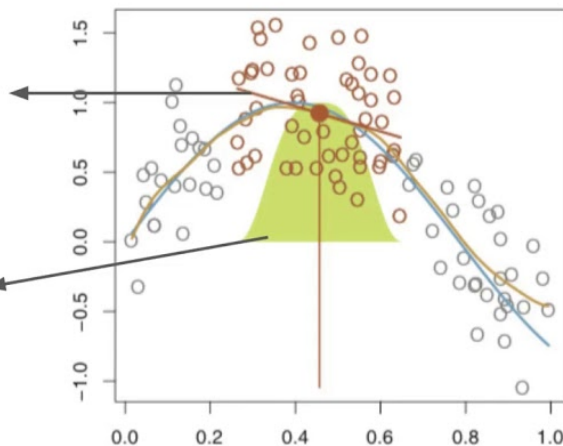


Figure 7.9 (ISLR)

What does linear mean

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


The estimation is a linear function in Y

What does linear mean

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

The estimation is a linear function in Y

Linear regression over a b c

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

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

“Feature extraction”
Lecture 15

a, b, c is linear in y

All quadratic function forms a (linear) vector space!

What does linear mean

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

The estimation is a linear function in Y

Linear regression over a b c

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

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

a, b, c is linear in y

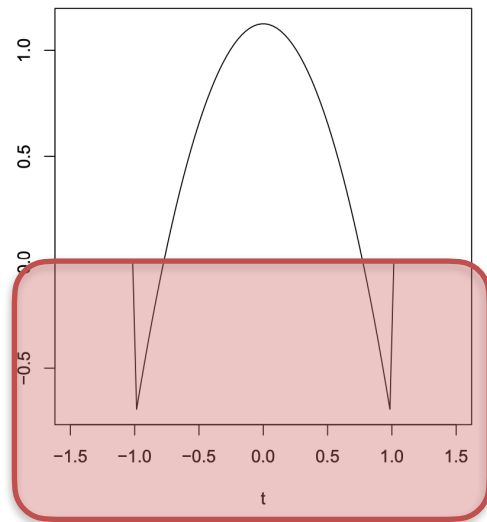
Linear smoothing =
local poly regression



All quadratic function forms a (linear) vector space!

Higher-order Kernel

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



Kernel Density Estimation

Let X_1, X_2, \dots, 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:

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

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

Bias

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

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

for some c .

Proof. We have

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

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 \hat{p}_h satisfies:*

$$\sup_{p \in \Sigma(\beta, L)} \text{Var}(\hat{p}_h(x)) \leq \frac{c}{nh^d} \quad (15)$$

for some $c > 0$.

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

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

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

Final Result

The optimal bound one can get

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

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



Estimating the derivatives

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

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

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

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

For any x_0 , and prove the MSE bound

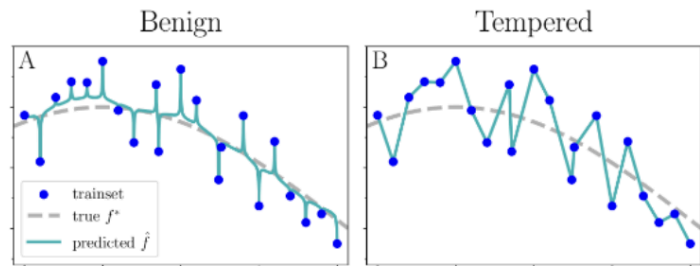
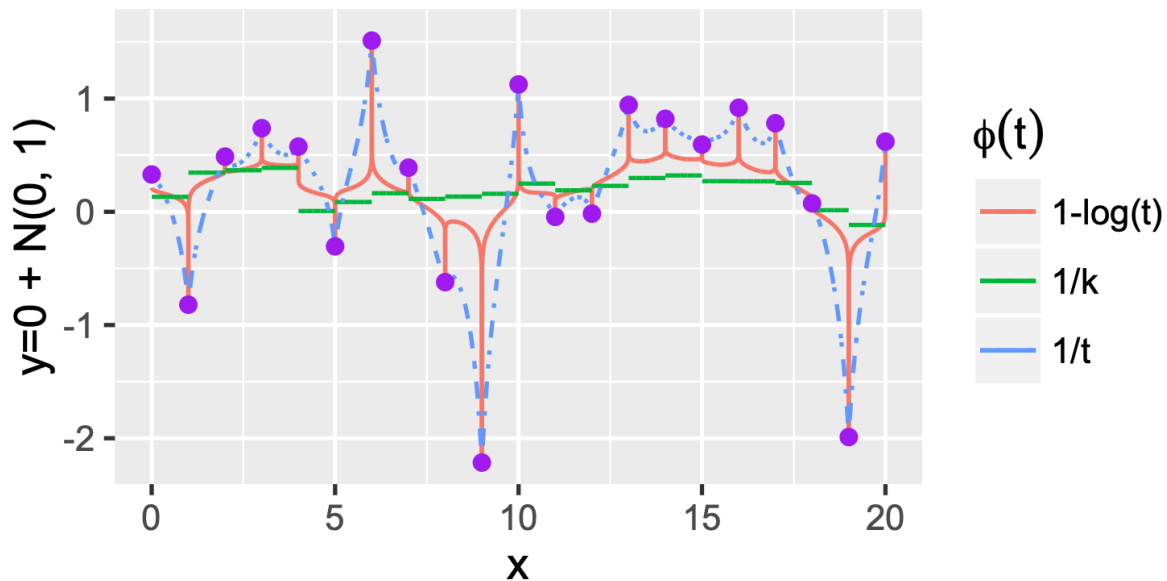
$$\mathbb{E}[|\hat{d}_n(x_0) - p'(x_0)|^2] \leq n^{-\frac{2(\beta-1)}{1+2\beta}}.$$

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.