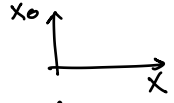
STAT 535

10/5/23

Lecture 4

Kernel predictors



- ☐ HW1
- ☐ Tutorial
  some of the predictors
- ☐ HW1 due 10/12 or
  HW2 due 10/19
- ☐ Quiz 1 10/19
  LI

Lecture Notes I – Examples of Predictors

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

September 25, 2023

Prediction problems by the type of output ✓

The "learning" paradigm and vocabulary ✓

Some concepts in Classification ✓

The Nearest-Neighbor and kernel predictors ✓ ←

Linear predictors

- Least squares regression
- Linear Discriminant Analysis (LDA)
- QDA (Quadratic Discriminant Analysis)
- Logistic Regression
- The PERCEPTRON algorithm

} Tutorial

Classification and regression tree(s) (CART)

The Naive Bayes classifier

Reading HTF Ch.: 2.3.1 Linear regression, 2.3.2 Nearest neighbor, 4.1–4 Linear classification, 6.1–3. Kernel regression, 6.6.2 kernel classifiers, 6.6.3 Naive Bayes, 9.2 CART, 11.3 Neural networks, Murphy Ch.: 1.4.2 nearest neighbors, 1.4.4 linear regression, 1.4.5 logistic regression, 3.5 and 10.2.1 Naive Bayes,4.2.1–3 linear and quadratic discriminant, 14.7.3– kernel regression, locally weighted regression, 16.2.1–4 CART, (16.5 neural nets), Bach Ch.:

Kernel regression and classification

- Like the K-nearest neighbor but with "smoothed" neighborhoods
- The predictor

$$f(x) = \sum_{i=1}^{n} \beta_i b(x, x^i) y^i \tag{6}$$

kernel

query point

coefficients

where β_i are coefficients

$b : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$

kernel $\approx$ similarity (x, x^i)

Ex: Gaussian kernel

$$b(x, x') = \frac{1}{(\sqrt{2\pi})^d} e^{-\frac{1}{2}\|x - x'\|^2}$$

Euclidean distance

Kernel regression and classification

- Like the K-nearest neighbor but with "smoothed" neighborhoods
- The predictor

$$f(x) = \sum_{i=1}^{n} \beta_i b(x, x^i) y^i \tag{6}$$

  where β_i are coefficients
- Intuition: center a "bell-shaped" *kernel* function b on each data point, and obtain the prediction $f(x)$ as a weighted sum of the values y^i, where the weights are $\beta_i b(x, x^i)$
- Requirements for a kernel function $b(x, x')$
  1. non-negativity
  2. symmetry in the arguments x, x'
  3. optional: radial symmetry, bounded support, smoothness
- A typical kernel function is the **Gaussian kernel** (or **Radial Basis Function (RBF)**)

$$b(z) \propto e^{-z^2/2} \tag{7}$$

standard

$$b_h(x, x') \propto e^{-\frac{||x-x'||^2}{2h^2}} \quad \text{with} \quad h = \textbf{the kernel width} \tag{8}$$

gaussian with stdev = h

Regression example

A special case in wide use is the Nadaraya-Watson regressor

$$\sum_{i=1}^{n} w_i y^i = f(x) = \frac{\sum_{i=1}^{n} b_h\left(\frac{||x-x^i||}{h}\right) y^i}{\sum_{i=1}^{n} b_h\left(\frac{||x-x^i||}{h}\right)}.$$

independent of i'

$$\beta_{i'} = \frac{1}{\sum_{i=1}^{n} b\left(\frac{||x-x^i||}{h}\right)} \quad (9)$$

In this regressor, $f(x)$ is always a convex combination of the y^i's, and the weigths are proportional to $b_h(x, x^i)$.
The Nadaraya-Watson regressor is biased if the density of P_X varies around x.

normalization

$$w_i = \frac{b(x, x^i)}{\sum_{i'=1}^{n} b(x, x^i)} \Rightarrow \sum_{i=1}^{n} w_i = 1$$

$$\Rightarrow \hat{y} = f(x) \leq \max y^i \geq \min y^i$$

Types of kernels

- Requirements for $b(\cdot,\cdot)$

⇓

Requirements for $b(z)$

1. $b(z) \geq 0$ for all z
2. $\int_{-\infty}^{\infty} b(z)\,dz = 1$

- $b(x,x') = b(\|x-x'\|)$

$b(z)$ function of distance only

$b : \mathbb{R} \to [0,\infty)$

Typical properties

3. $b(z) = b(-z)$ symmetry
4. $b(z) \searrow$ for $z \geq 0$ decreasing $\Rightarrow \max_z b(z) = b(0)$

Support of b

$$\text{supp } b = \begin{cases} [-a,a] & a>0 \text{ finite (compact)} \\ \mathbb{R} & \text{infinite supp} \end{cases}$$

Ex: Gaussian b

Smoothness

b', b'' exist almost everywhere

$$\text{supp } g = \{x \mid g(x) \neq 0\}$$

$g : \mathbb{R}^d \to \mathbb{R}$

if $g \geq 0 \Rightarrow$

$$\text{supp } g = \{x \mid g(x) > 0\}$$

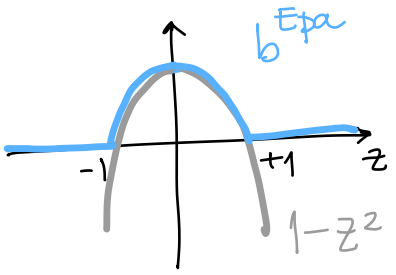
Epanechnikov kernel

$$b(z) = c\left(1 - z^2\right)_+$$

↑

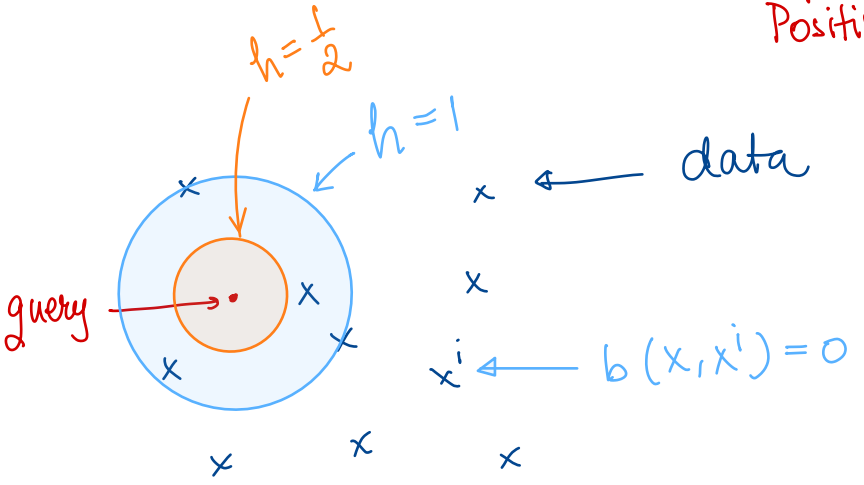
quadratic

finite support



$$g_+(x) = \begin{cases} g(x) & g(x) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

↑ Positive part of g



$h = \frac{1}{2}$

Kernel width (Bandwidth) h

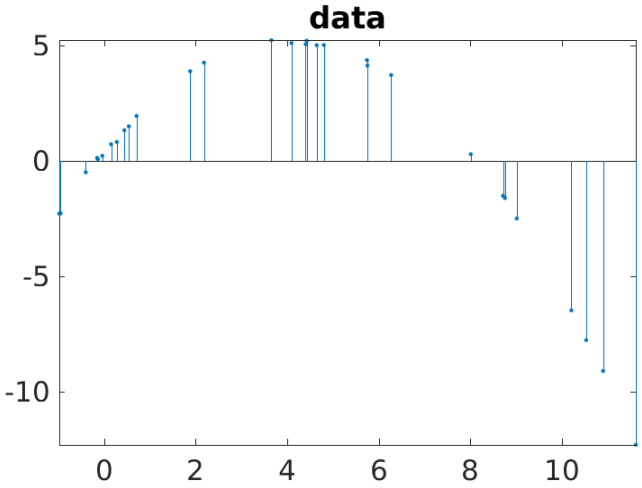
For any $b(z)$ kernel

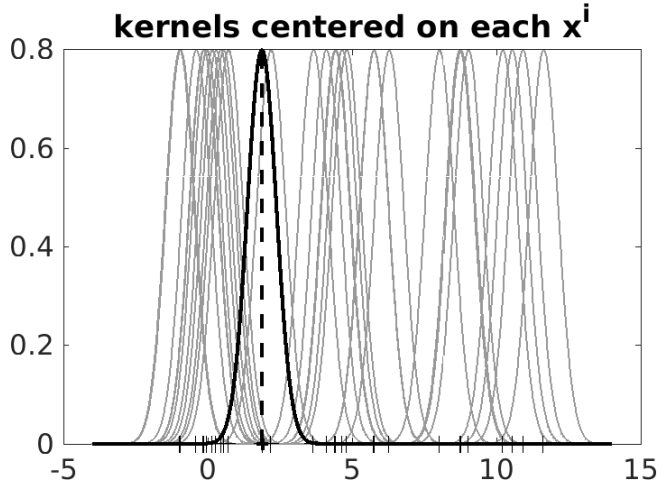
$$b_h(z) = \frac{1}{h} b\left(\frac{z}{h}\right)$$ ← check this!

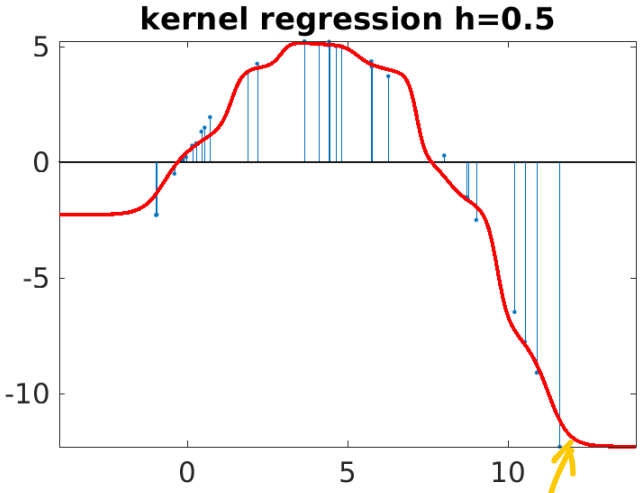
- h is a smoothing parameter

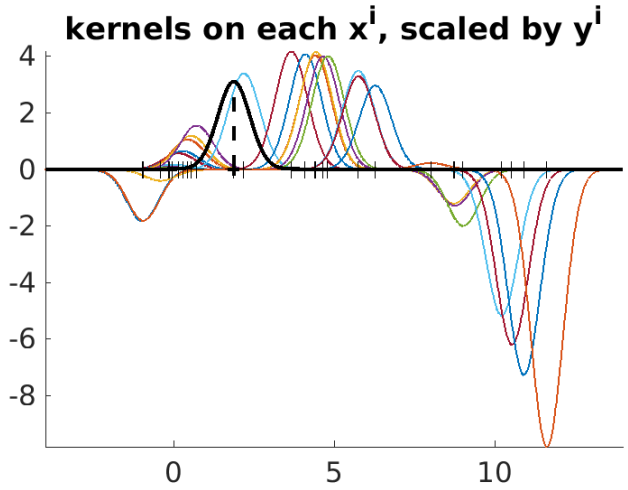
$h \uparrow \Rightarrow f$ smoother

An example: noisy data from a parabola









bias!
Removed by Local Linear Regression

Local Linear Regression

To correct for the bias (to first order) one can estimate a regression line around x.

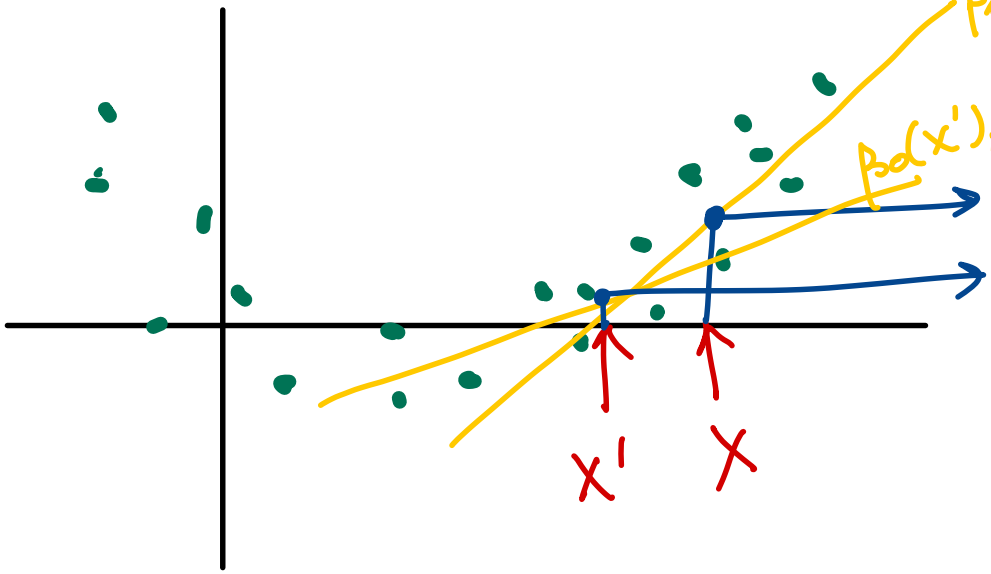
1. Given **query point** x
2. Compute kernel $b_h(x, x^i) = w_i$ for all $i = 1, \ldots N$
3. Solve weighted regression $\min_{\beta,\beta_0} \sum_{i=1}^{N} w_i \left(y^i - \beta^T x^i - \beta_0\right)^2$ to obtain β, β_0 (β, β_0 depend on x through w_i)
4. Calculate $f(x) = \beta^T x + \beta_0$

↳ local y^i's on local x^i's

Exercise Show that Nadaraya-Watson solves a local linear regression with fixed $\beta = 0$

b_h, x $\rightarrow$ $w_i(x) \propto b_h(x, x^i)$

↳ weight of x^i in predicting x



Kernel binary classifiers

- Obtained from Nadaraya-Watson by setting y^i to ± 1.
- Note that the classifier can be written as the difference of two non-negative functions

$$f(x) \propto \sum_{i:y^i=1} b\left(\frac{||x - x^i||}{h}\right) - \sum_{i:y^i=-1} b\left(\frac{||x - x^i||}{h}\right). \tag{10}$$

$$f(x) = \sum_{i:y^i=1} w_i(x) + (-1) \sum_{i:y^i=-1} w_i(x)$$

$$\hat{y}(x) = \operatorname{sgn} f(x)$$

$$f(x) \in [-1, 1]$$

$$f_{(x)} \propto g_{(x)} \iff$$

$$f_{(x)} = c\, g_{(x)} \text{ for } c \in \mathbb{R} \text{ constant}$$