Linear Smoothers II

`numpy` 101; binary classification; kernel smoothers; density estimation

DATA 607 — Session 2 — 27/02/2019

Session Plan

1. `numpy` 101
   - axes; universal functions; broadcasting
   - exercise
2. nonparametric binary classification
   - decision functions
   - zero-one loss; misclassification rate
   - examples/demos (`sklearn`)
3. kernel regression and classification
   - kernel functions; examples
   - Nadaraya-Watson smoothers
   - examples/demos (`sklearn`)
4. density estimation
   - histogram estimators
   - kernel estimators
   - connection with kernel regression and classification
5. summatory exercise

1. numpy 101

numpy_101.ipynb

Regression with categorical response

Setting:

- two *classes* with *labels* 0 and 1
- (X, Y) jointly distributed, $X \in \mathbb{R}^n$, $Y \in \{0, 1\}$
- $L(\widehat{y}, y)$ a *loss function*

Goal: Find a function

$$\widehat{r} : \mathbb{R}^n \longrightarrow \mathbb{R}$$

that minimizes the average risk:

$$\mathbb{E}\big[L(\widehat{r}(X), Y)\big]$$

$L(\widehat{f}(X), Y)$ is the penalty for predicting $\widehat{f}(X)$ when the target was Y.

Example: Naïve Bayes Classifier

Let L is *zero-one loss*:

$$L(\widehat{y}, y) = \begin{cases} 0 & \text{if } \widehat{y} = y, \\ 1 & \text{if } \widehat{y} \neq y \end{cases}$$

The average risk, $\mathbb{E}\big[L(\widehat{r}(X), Y)\big]$, is minimized by the *Bayes estimator*:

$$\widehat{f}(x) := \underset{y \in \{0,1\}}{\operatorname{argmax}} P(X = x | Y = y) P(Y = y)$$

This is not so useful if we don't know $P(x, y)$!

Suppose K is odd.

Let $\widehat{r}$ be the nearest neighbors regressor associated to the dataset

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Since $Y_i \in \{0, 1\}$, we have $\widehat{r}(x) \in \left\{\frac{0}{K}, \frac{1}{K}, \ldots, \frac{K}{K}\right\}$ for all x.

Let

$$\widehat{f}(x) := \widehat{r}(x), \text{ rounded to the nearest integer.}$$

Let

$$K_k(x, x_i) = \begin{cases} 1 & \text{if } x_i \text{ is one of the } k\text{-nearest neighbors of } x \\ 0 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} \widehat{r}_k(\vec{x}) &= \frac{1}{k} \sum_{i=1}^{n} K_k(x, x_i) Y_i \\ &= \frac{\sum_{i=1}^{n} K_k(x, x_i) Y_i}{\sum_{i=1}^{n} K(x, x_i)} \end{aligned}$$

Let

$$K_h(x, x_i) = \begin{cases} 1 & \text{if } \|x - x_i\| \leq h \\ 0 & \text{otherwise.} \end{cases}$$

$$\widehat{r}_h(x) = \frac{\sum_{i=1}^n K_h(x, x_i) Y_i}{\sum_{i=1}^n K_h(x, x_i)}$$

If

$$K(x) = \mathbf{1}_{[-1,1]}(x) = \begin{cases} 1 & \text{if } x \in [-1, 1], \\ 0 & \text{otherwise,} \end{cases}$$

then

$$K_h(x, x_i) = K\left(\frac{x - x_i}{h}\right)$$

and

$$\widehat{r}_h(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}$$

Definition

A *kernel function* is a function $K(x)$ such that

1. $K(x) \geq 0$,
2. $K(-x) = K(x)$,
3. $\int_{-\infty}^{\infty} K(x)dx = 1$,

Definition

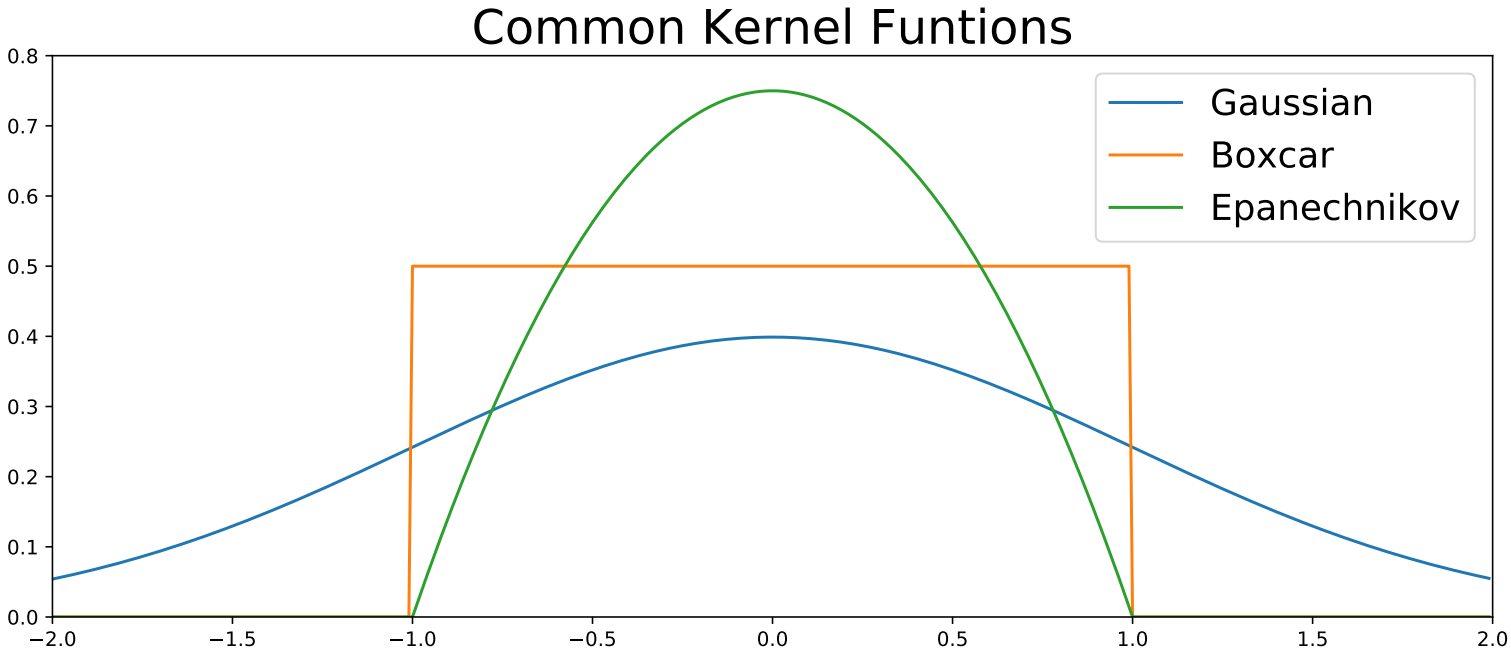
The *Nadaraya-Watson smoother* with bandwidth $h > 0$ associated to the kernel function $K(x)$ and the dataset $(x_1, Y_1), \ldots, (x_n, Y_n)$ is

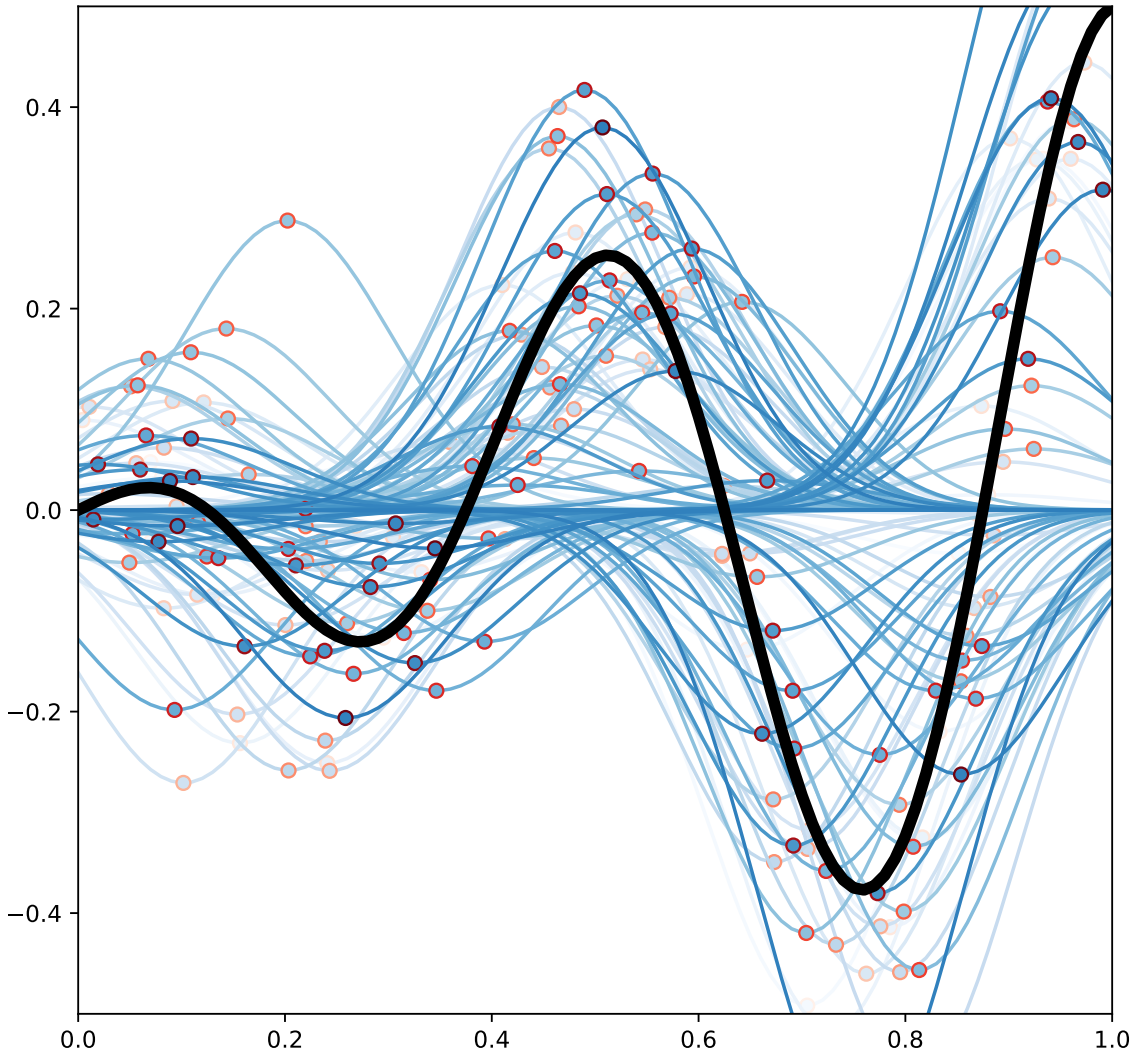
$$\widehat{r}_h(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

Examples of kernels

1. The *boxcar kernel*:
$$K(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$$
2. The *Gaussian kernel*:
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$
3. The *Epanechnikov kernel*:
$$K(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{[-1,1]}(x)$$

The Nadaraya-Watson smoother associated to the boxcar kernel is the local average smoother.


Common Kernel Funtions
Gaussian
Boxcar
Epanechnikov
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0

You can use kernels to estimate probability density functions.

Definition

The *kernel density estimator* associated to a kernel function $K(x)$, a bandwidth $h > 0$, and a dataset $X_1, \ldots, X_n$ is

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

To determine the optimal h, minimize the cross validation score.

Definition

The *leave one out cross validation score* associated to a kernel density estimator $\widehat{f}_n$ with bandwidth h is

$$\widehat{R}(h) := \int \widehat{f}_n(x)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_n^{(-i)}(X_i)$$