DATA 607 Statistical and Machine Learning

Session 3: Kernel Smoothers; Nonparametric Classifiers

Matthew Greenberg

Department of Mathematics and Statistics
University of Calgary

22.03.2020

This Evening's Agenda

1. Kernel Smoothers
   - Weighted Averages
   - The Boxcar Kernel Smoother
   - Kernel Functions
   - Kernel Smoothers
   - Examples

2. Classfication

Weighted Averages

Suppose I'm computing the course grade for my MATH 307 (Complex Analysis I) students.

Their course grades (G) are computed based on three assignments (A1, A2, A3), two tests (T1, T2), and a final exam (F).

In the course grade computation, a midterm is assigned twice the weight of an assignment and the final exam is assigned twice the weight of a midterm.

Fill in the grade column. (All scores are percentages.)

Student	**A1**	**A2**	**A3**	**T1**	**T2**	**F**	**G**
James	70	80	50	75	80	40	
Anton	60	90	95	70	90	85	
Hiraku	90	95	100	95	95	100	

The course grades are **weighted averages** of the students' scores on the individual course components.

$$\mathrm{G} = \frac{1 \cdot \mathrm{A1} + 1 \cdot \mathrm{A2} + 1 \cdot \mathrm{A3} + 2 \cdot \mathrm{T1} + 2 \cdot \mathrm{T1} + 4 \cdot \mathrm{F}}{1+1+1+2+2+4}$$

James:

$$G = \frac{1 \cdot 70 + 1 \cdot 80 + 1 \cdot 50 + 2 \cdot 75 + 2 \cdot 75 + 4 \cdot 40}{1+1+1+2+2+4} = 60.00$$

Anton:

$$G = \frac{1 \cdot 60 + 1 \cdot 90 + 1 \cdot 95 + 2 \cdot 70 + 2 \cdot 90 + 4 \cdot 85}{1+1+1+2+2+4} = 82.27$$

Hiraku:

$$G = \frac{1 \cdot 90 + 1 \cdot 95 + 1 \cdot 100 + 2 \cdot 95 + 2 \cdot 95 + 4 \cdot 100}{1+1+1+2+2+4} = 96.81$$

Given **weights** $w_1, \ldots, w_n$, the associated The **weighted average** of $x_1, \ldots, x_n$ is

$$\frac{\sum_i w_i x_i}{\sum_i w_i} = \frac{w_1 x_1 + \cdots + w_n x_n}{w_1 + \cdots + w_n}.$$

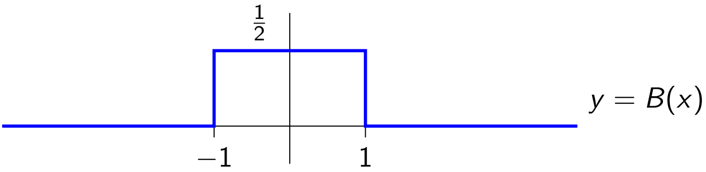
When the weights are all equal, say $w_1 = \cdots = w_n = w$, the weighted average is just the usual average:

$$\frac{\sum_i w x_i}{\sum_i w} = \frac{w \sum_i x_i}{nw} = \frac{\sum_i x_i}{n}$$

Boxcar Kernel

Define the **Boxcar Kernel** by

$$
\begin{aligned}
B(x) &= \frac{1}{2}\mathbf{1}_{(-1,1)}(x) \\
&= \begin{cases} \frac{1}{2} & \text{if } -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases} \qquad (\mathbf{x} \in \mathbb{R}).
\end{aligned}
$$



Boxcar Kernel Smoother

Data set:

$$\mathcal{D} = \{(\mathbf{x}_1, y_1),\ (\mathbf{x}_1, y_1),\ \ldots,\ (\mathbf{x}_n, y_n)\}$$

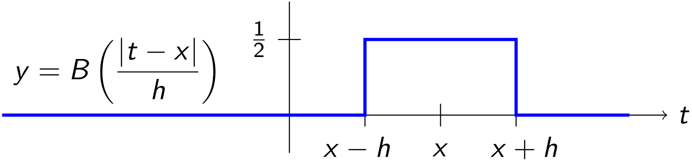
Boxcar Kernel Smoother:

$$\widehat{r}(\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i \, B\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{h}\right)}{\sum_{i=1}^{n} B\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{h}\right)}$$

This is a **weighted average** of the y_i. All y_i for which $\mathbf{x}_i$ is within a distance h of $\mathbf{x}$ are assigned weight $w_i = 1$; all others are assigned weight $w_i = 0$.

$$B\left(\frac{\|x_i - x\|}{h}\right) = \begin{cases} 1 & \text{if } \|x_i - x\| < h \\ 0 & \text{otherwise} \end{cases}$$

$$y = B\left(\frac{|t - x|}{h}\right)$$



Note: Boxcar Kernel Smoother = Sliding Window Smoother

Generalization: Replace B with another “kernel” function.

Kernel Functions

$K(x)$ is a **kernel function** if

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

Popular Kernels

1. Boxcar:

$$B(x) = \frac{1}{2}\mathbf{1}_{(-1,1)}(x)$$

2. Triangular:

$$T(x) = (1 - |x|)\mathbf{1}_{(-1,1)}(x)$$

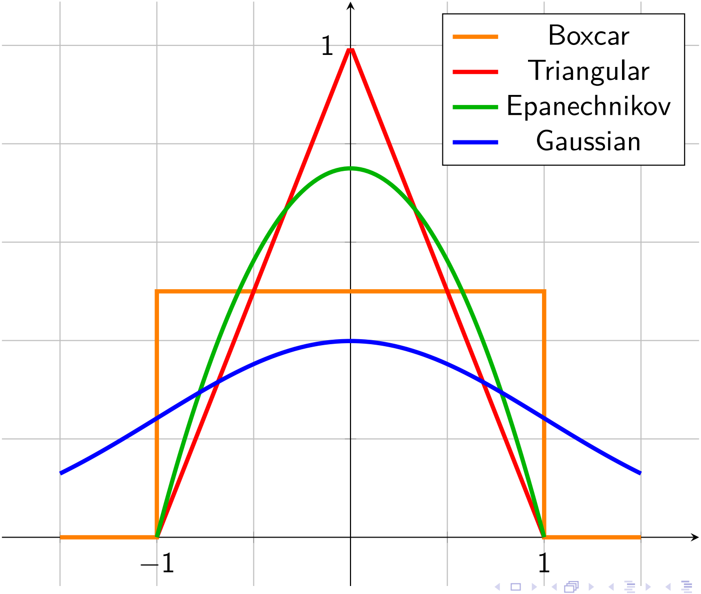
3. Epanechnikov:

$$E(x) = \frac{3}{4}(1 - x^2)\mathbf{1}_{(-1,1)}(x)$$

4. Gaussian:

$$G(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

Popular Kernels



Kernel Smoothers

Definition

The **kernel smoother** associated to the data set

$$\mathcal{D} = \{(\mathbf{x}_1, y_1),\ (\mathbf{x}_1, y_1),\ \ldots,\ (\mathbf{x}_n, y_n)\},$$

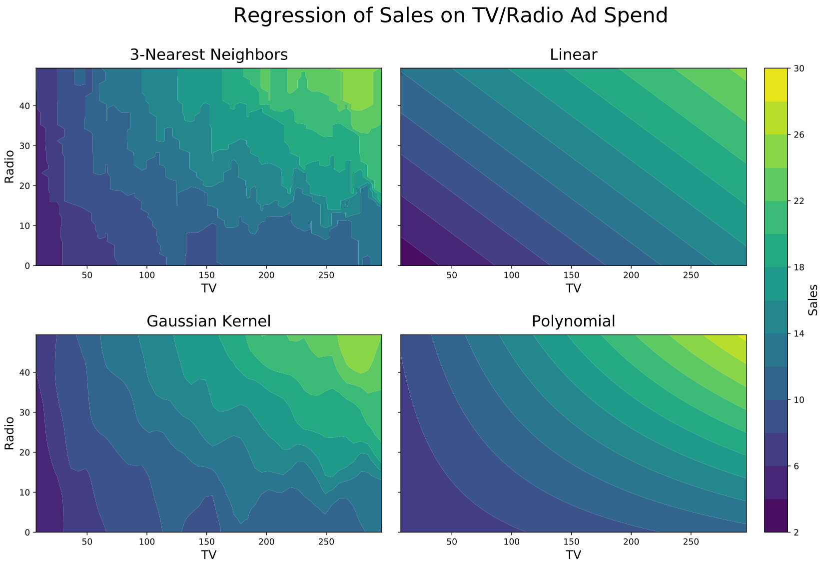
a kernel function K, and a bandwidth $h > 0$ is the function $\widehat{r}$ defined by

$$\widehat{r}(\mathbf{x}) = \frac{\displaystyle\sum_{i=1}^{n} y_i\, K\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{h}\right)}$$

This is a **weighted average** of $y_1, \ldots, y_n$ with y_i having weight

$$w_i = K\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{h}\right).$$

Comparison of Regression Models



Model	Training MSE	Testing MSE
Linear Regression	2.10	5.78
Polynomial Regression	0.51	2.51
k-Nearest Neighbors ($k = 3$)	0.49	1.82
Gaussian Kernel ($h = 6$)	0.21	1.13

- `advertising.csv`
- TV, Radio, and Sales columns only
- 160 training samples, 40 testing samples

Binary Classification

- Just like regression, but targets in $\{0, 1\}$ instead of $\mathbb{R}$.
- (X, Y) jointly distributed, $X \in \mathbb{R}^n$, $Y \in \{0, 1\}$
- $r(X) = \mathbb{E}[Y|X] = P[Y = 1|X] \in [0, 1]$
- Bayes Estimator:

$$\widehat{r}_{\text{Bayes}} = \begin{cases} 1 & \text{if } P[Y = 1|X] \geq 1/2, \\ 0 & \text{otherwise.} \end{cases}$$