CSci5521: Machine Learning Fundamentals

- Nonparametric Methods

Catherine Qi Zhao
Computer Science and Engineering
University of Minnesota

Parametric Method

- Assume that the sample is drawn from some distribution that obeys a known model (e.g., Bernoulli, Gaussian)
- The model is defined up to a number of parameters
- Learning is to fit the model with the best parameters to the data

Parametric Estimation

- Types: density estimation, classification, regression
- Advantages
- Disadvantages

Nonparametric Estimation

- All we assume: Similar inputs have similar outputs
- Functions (pdf, discriminant, regression) change smoothly
- No single global model, but local models: Given x, find a small number of closest training instances and interpolate from these
- Aka lazy/memory-based/case-based/instance-based learning

Nonparametric Estimation

- Types: density estimation, classification, regression
- Advantages
- Disadvantages

Notations

N data points

- Sample set: $X = \{x^t\}_{t=1}^{N}$

PDF

- Probability density function: $p(x)$
- Estimator of the probability density function: $\hat{p}(x)$
- Cumulative density function: $F(x)$
- Estimator of the cumulative density function: $\hat{F}(x)$
- Window/interval/bin/neighborhood size: h

Nonparametric Density Estimation

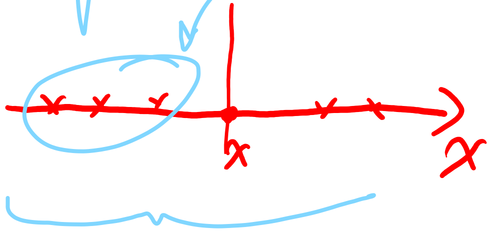
- The nonparametric estimator for the cumulative distribution function, *F(x)*, at point *x*

Cumulative Estimations

$$\hat{F}(x) = \frac{\#\{x^t \le x\}}{N}$$

\# of data points $x^t \le x$

Total

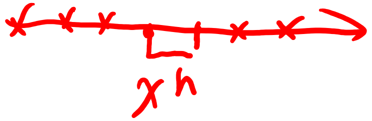


N

- The nonparametric estimate for the density function is

$\frac{d\hat{F}(x)}{dx}$

$$\hat{p}(x) = \frac{1}{h}\left[\frac{\#\{x^t \le x+h\} - \#\{x^t \le x\}}{N}\right]$$



Histogram Estimator

- Given the training set $X = \{x^t\}_{t=1}^N$ drawn iid from $p(x)$

- The bins are the intervals $[x_o + mh, x_o + (m+1)h]$
  origin x_o, bin size h

$[x_0, x_0+h)$
$[x_0+h, x_0+2h]$

- Histogram: $$\hat{p}(x) = \frac{\#\{x^t \text{ in the same bin as x}\}}{Nh}$$

num of data pts in same bin as x

bin size

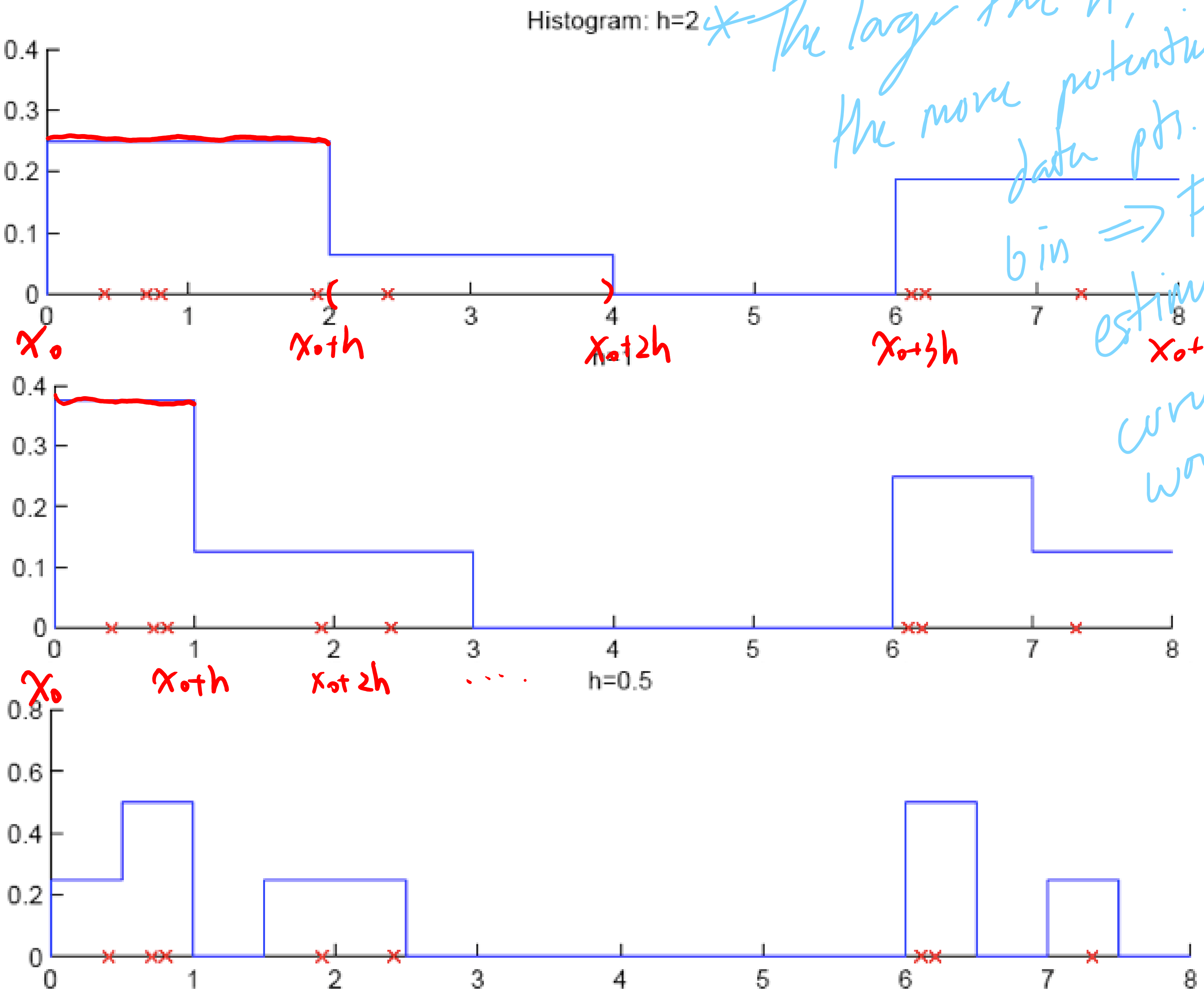
Histogram Estimator Example

- X = {0.4, 0.75, 0.8. 1.9, 2.4, 6.1, 6.2, 7.3}

$$\hat{p}(x) = \frac{\#\{x^t \text{ in the same bin as x}\}}{Nh}$$

$$\frac{4}{Nh = 8 \times 2} = 0.25$$

$$\frac{3}{Nh = 8 \times 1} = 0.375$$



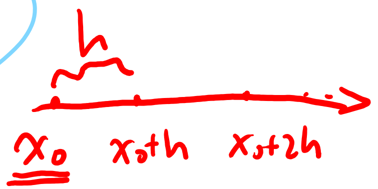
* The large the h, the more potential for data pts. in single bin ⇒ final estimation curve would be smoother.

Naive Estimator

fixed around @ x_0

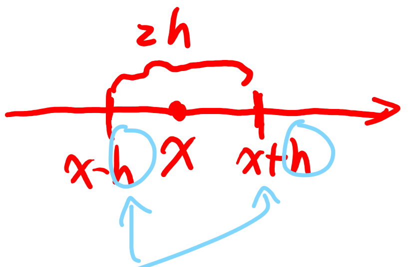
- Given the training set $X = \{x^t\}_{t=1}^N$ drawn iid from $p(x)$

- Histogram: $\hat{p}(x) = \dfrac{\#\{x^t \text{ in the same bin as x}\}}{Nh}$



from $x-h$ to $x+h$

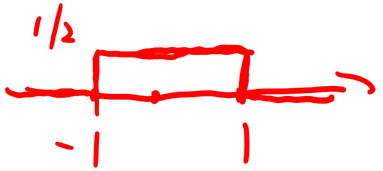
- Naive estimator: $\hat{p}(x) = \dfrac{\#\{x - h < x^t \leq x + h\}}{2Nh}$

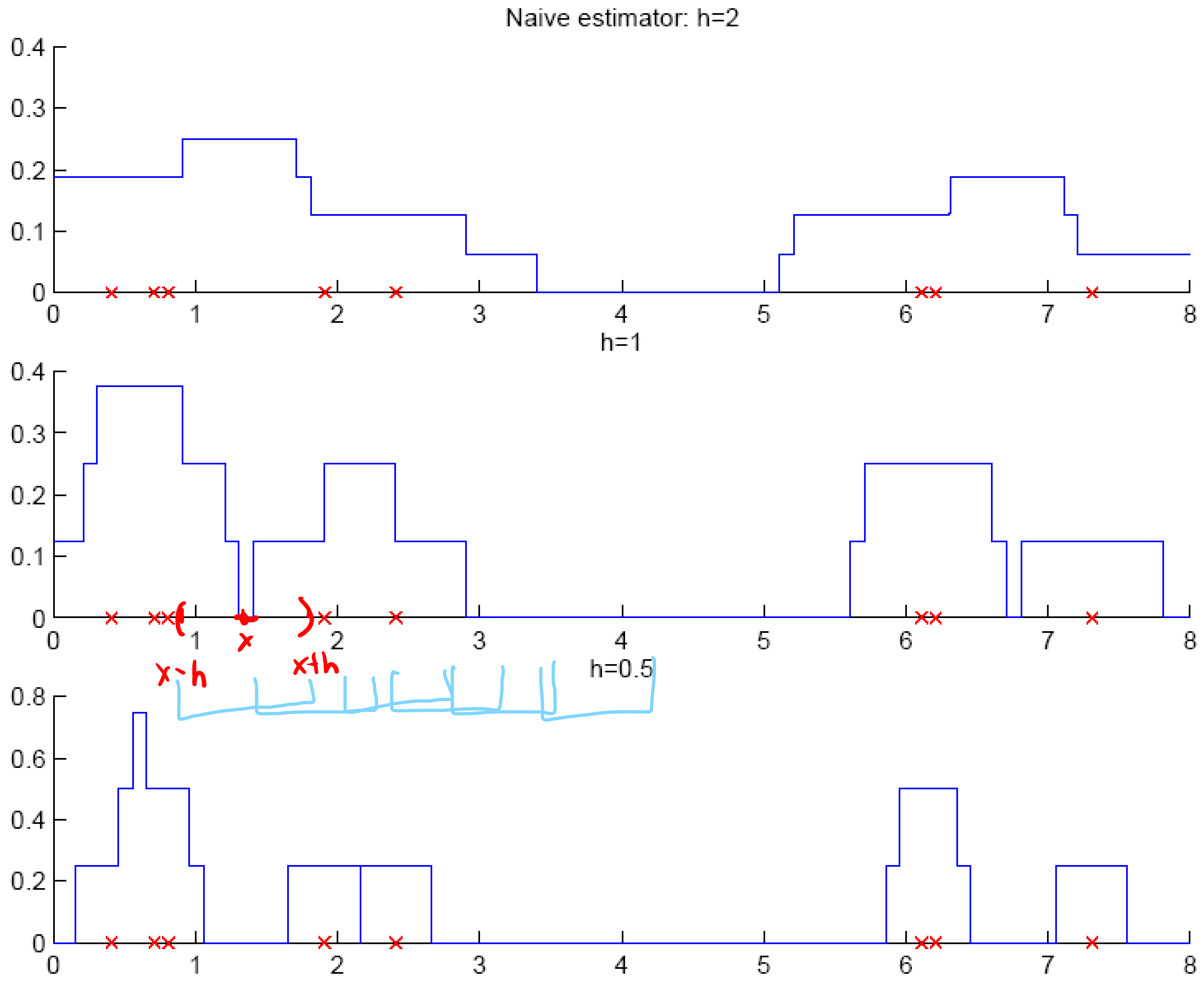


2 h's.

or

$$\hat{p}(x) = \frac{1}{Nh}\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right) \qquad w(u) = \begin{cases} \frac{1}{2}, & \text{if } |u| < 1 \\ 0, & \text{otherwise} \end{cases}$$



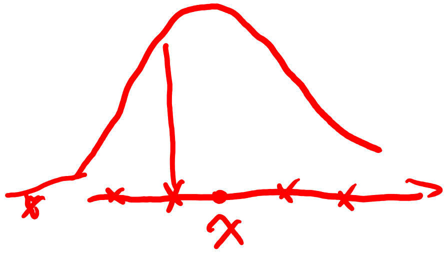


Naive estimator: h=2
h=1
h=0.5
x-h
x
x+h

Kernel Estimator

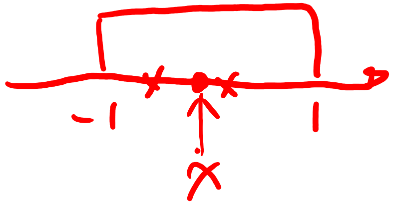
- Kernel function, e.g., Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{u^2}{2}\right]$$



- Kernel estimator (Parzen windows)

$$\hat{p}(x) = \frac{1}{Nh} \sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)$$

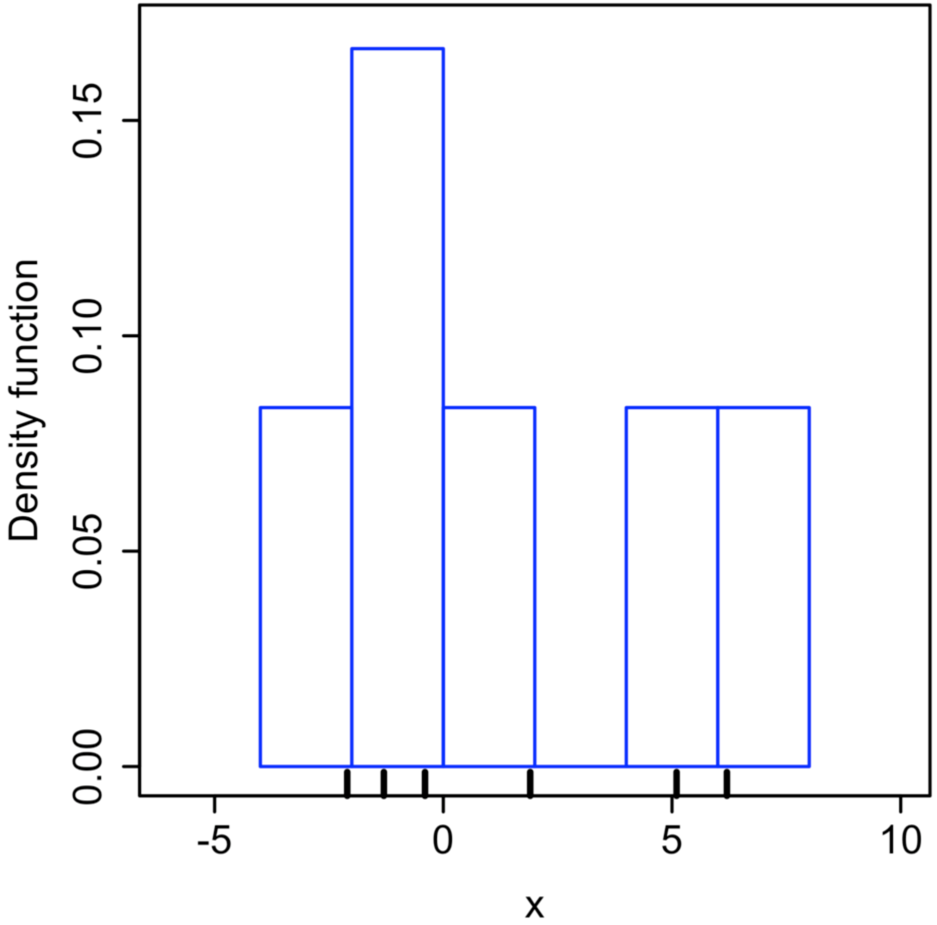


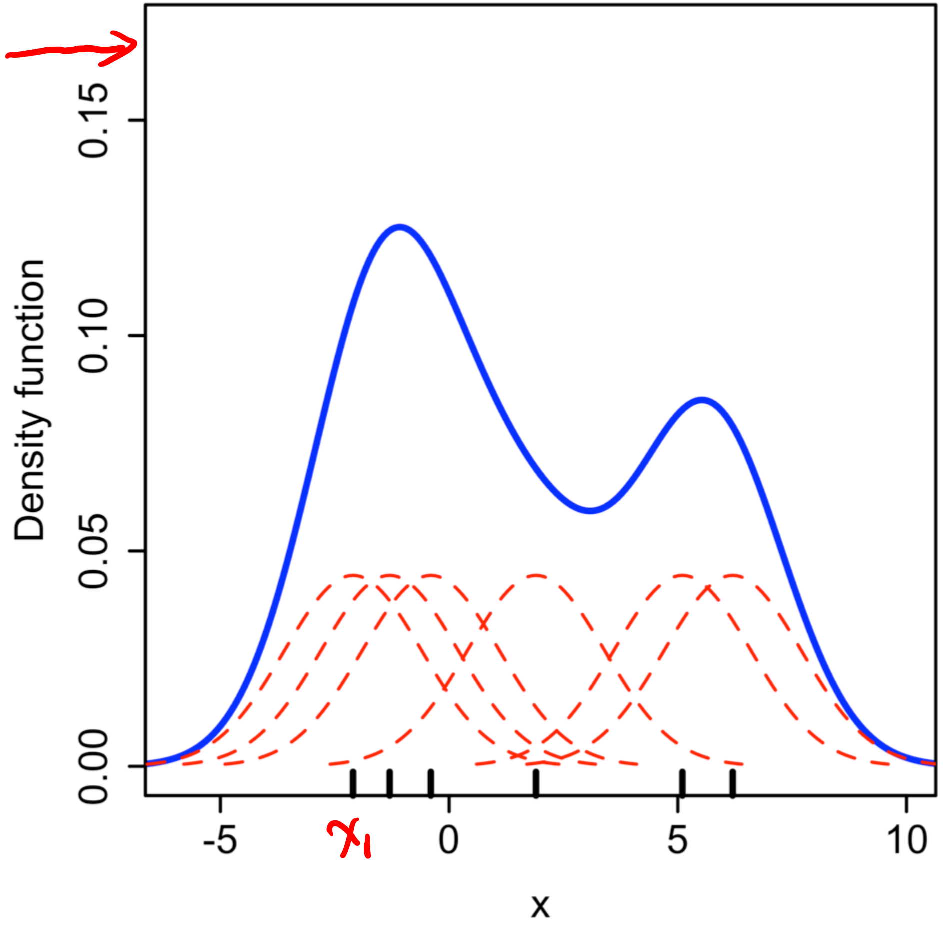
Kernel Estimator

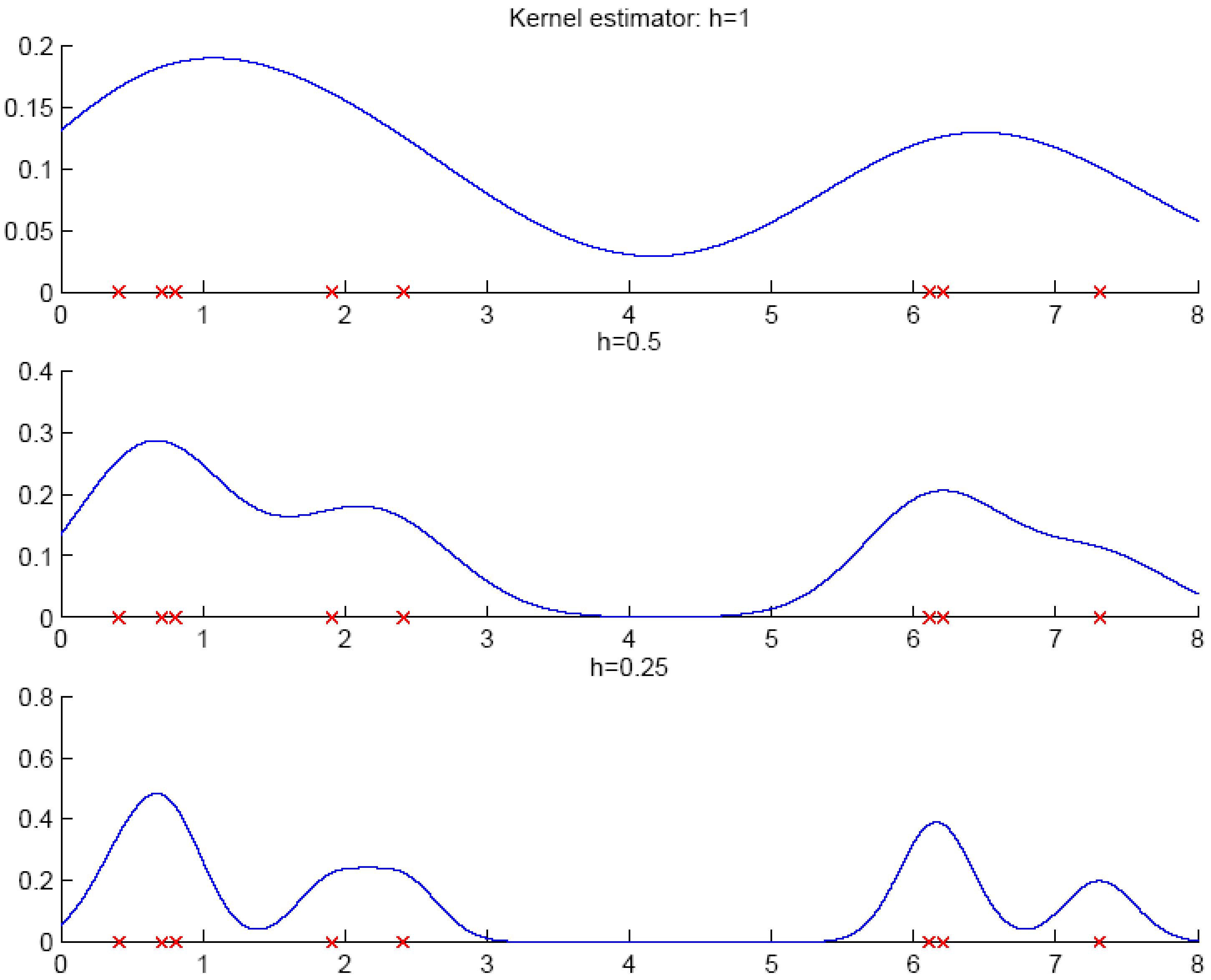
HISTOGRAM

KERNEL

- X = {-2.1, -1.3, -0.4, 1.9, 5.1, 6.2}






Kernel estimator: h=1
0.2
0.15
0.1
0.05
0
0 1 2 3 4 5 6 7 8
h=0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8
h=0.25
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8

k-Nearest Neighbor Estimator

- Instead of fixing bin width *h* and counting the number of instances, fix the instances (neighbors) *k* and check bin width

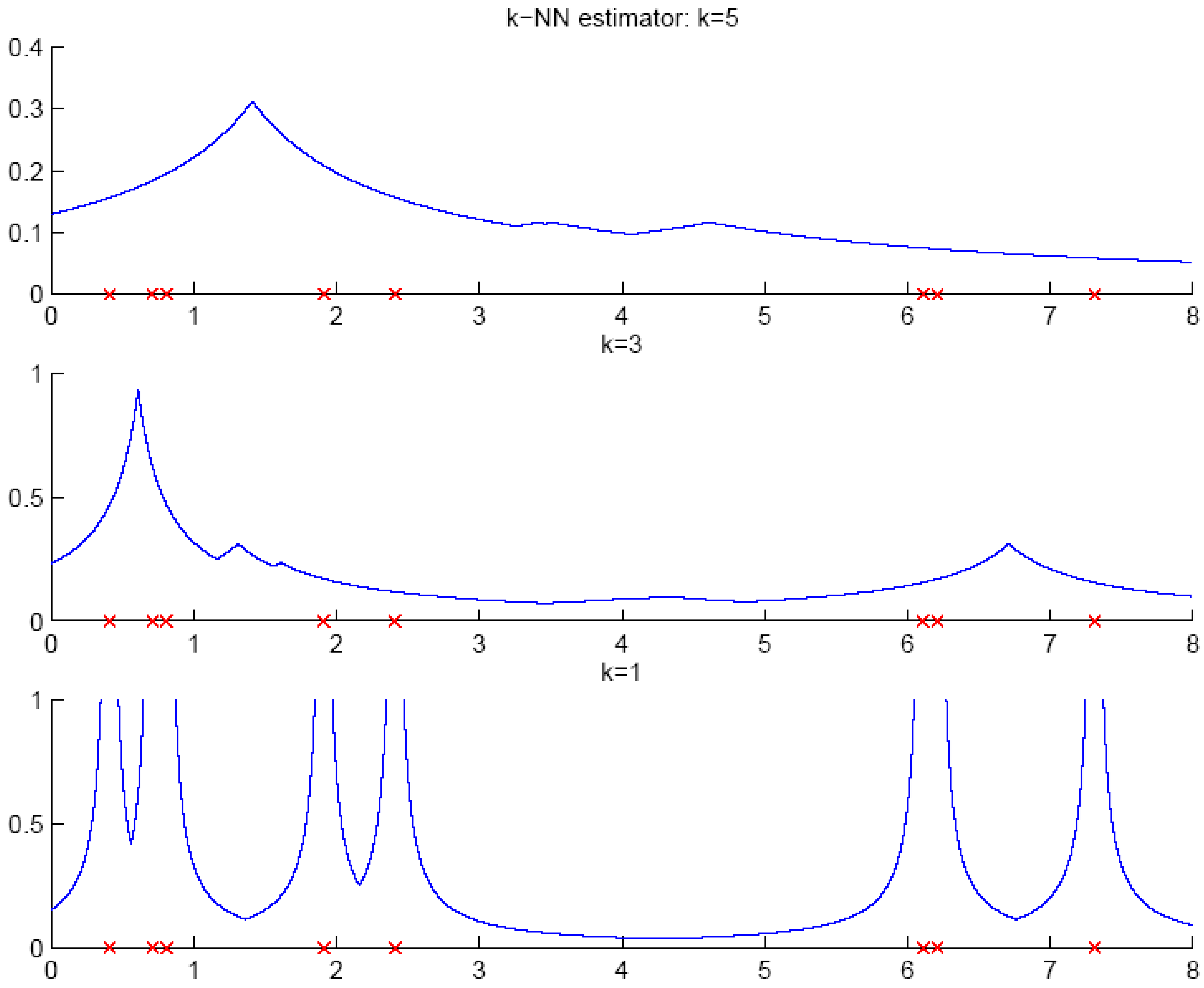
$$\hat{p}(x) = \frac{k}{2N\underbrace{d_k(x)}_{h}}$$

previously, fix h, k could vary

now, fix k, h could vary

$d_k(x)$, distance to *k*th closest instance to *x*

h


k−NN estimator: k=5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8
k=3
1
0.5
0
0 1 2 3 4 5 6 7 8
k=1
1
0.5
0
0 1 2 3 4 5 6 7 8

Nonparametric Classification – Kernel Estimator

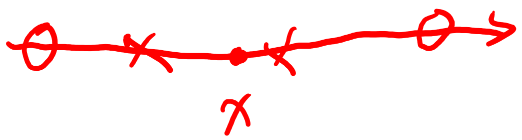
- Estimate $p(x|C_i)$ and use Bayes' rule

- Kernel estimator

non-parametric

$$\hat{p}(x|C_i) = \frac{1}{N_i h^d} \sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right) r_i^t \qquad \hat{P}(C_i) = \frac{N_i}{N}$$

$$g_i(x) = \hat{p}(x|C_i)\hat{P}(C_i) = \underline{\frac{1}{N h^d} \sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right) r_i^t}$$

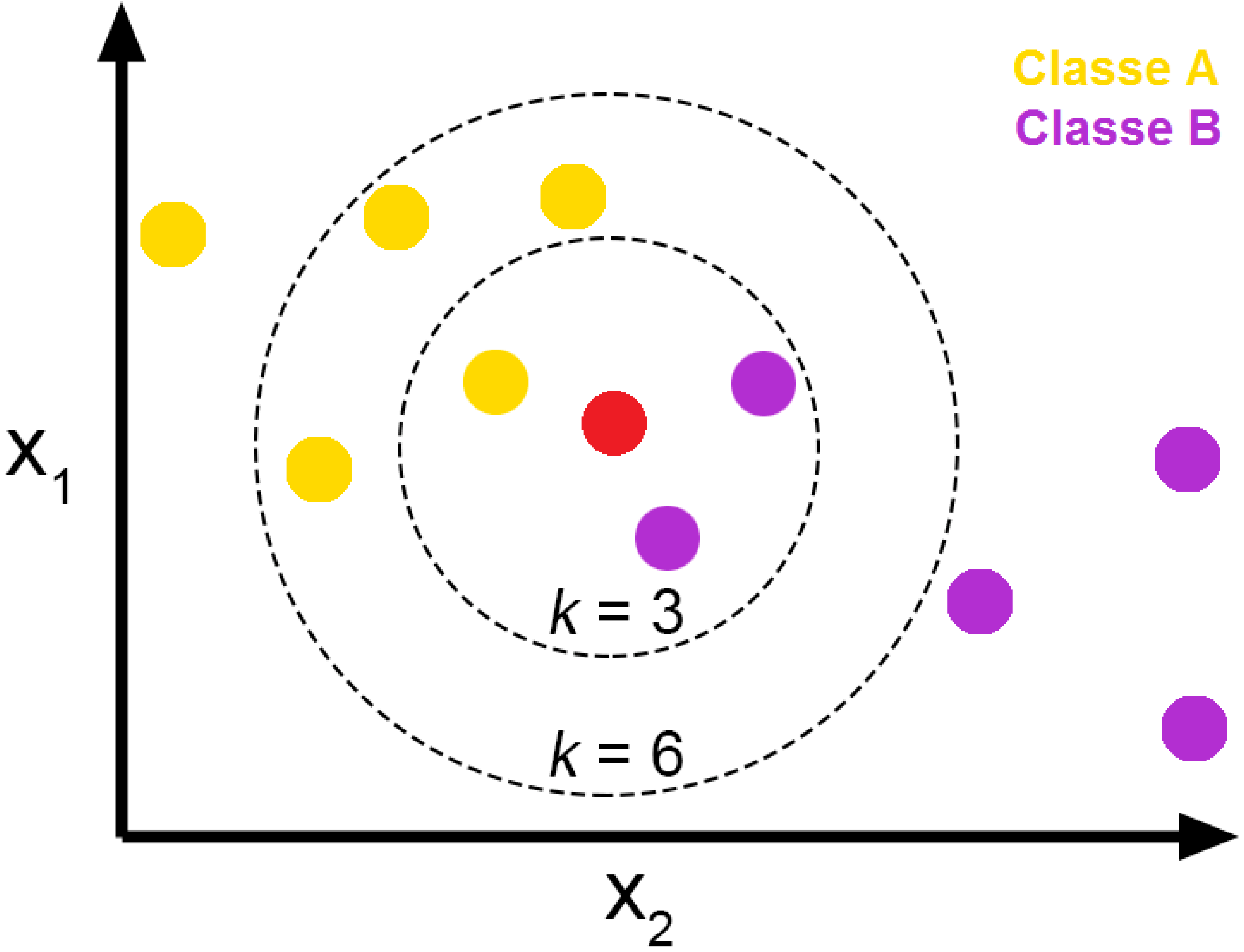


Nonparametric Classification – K-NN Estimator

- Estimate $p(x|C_i)$ and use Bayes' rule
- k-NN estimator

$$\hat{p}(x|C_i) = \frac{k_i}{N_i V^k(x)} \qquad \hat{P}(C_i|x) = \frac{\hat{p}(x|C_i)\hat{P}(C_i)}{\hat{p}(x)} = \frac{k_i}{k}$$

- When k = 1, nearest neighbor classifier



Classe A
Classe B
x_1
x_2
$k = 3$
$k = 6$

How to Choose k or h ?

- When k or h is small, single instances matter
- As k or h increases, we average over more instances
- Cross-validation can be used to finetune k or h.

Parametric vs Nonparametric

in general global models

in general local models

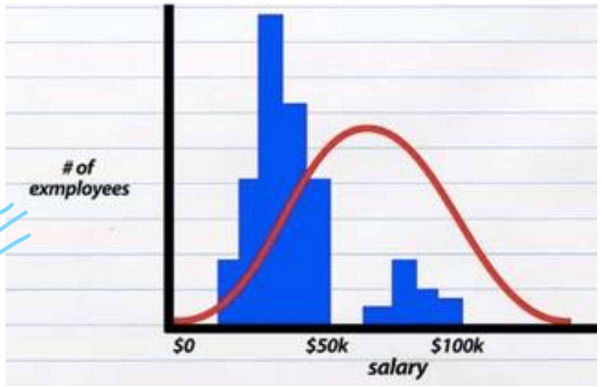
requires

stronger

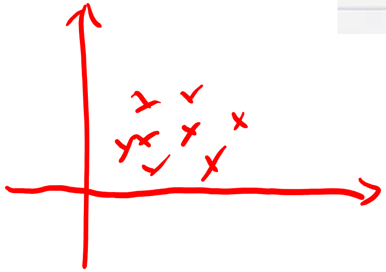
- Assumption
- Statistical power

less affected b/c local

more so



- Sensitivity to outliers



- Efficiency
- ......

Some materials credit to former 5521, Introduction to Machine Learning by Ethem Alpaydin and online resources