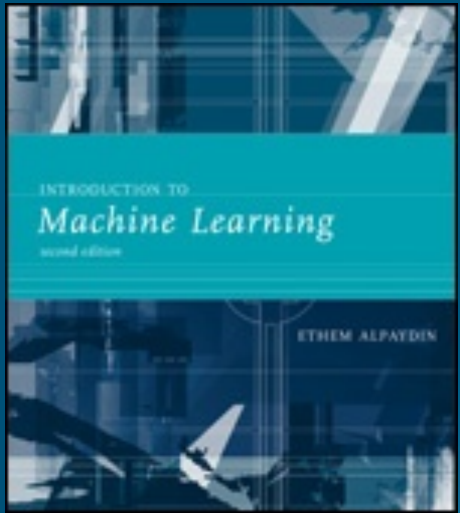


Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e

CHAPTER 8:

Nonparametric Methods

Nonparametric Estimation

- Parametric (single global model), semiparametric (small number of local models)
- Parametric: model parameters contain summary of the information in the dataset
- Nonparametric: Similar inputs have similar outputs
- Functions (pdf, discriminant, regression) change smoothly
- Keep the training data;"let the data speak for itself"
- Given x, find a small number of closest training instances and interpolate from these
- Aka lazy/memory-based/case-based/instance-based learning

Density Estimation

- Given the training set $X=\{x^t\}_t$ drawn iid from $p(x)$
- Divide data into bins of size h
- Histogram:

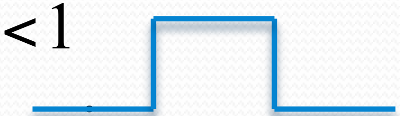
$$\hat{p}(x)=\frac{\#\{x^t \text{ in the same bin as } x\}}{Nh}$$

- Naive estimator:

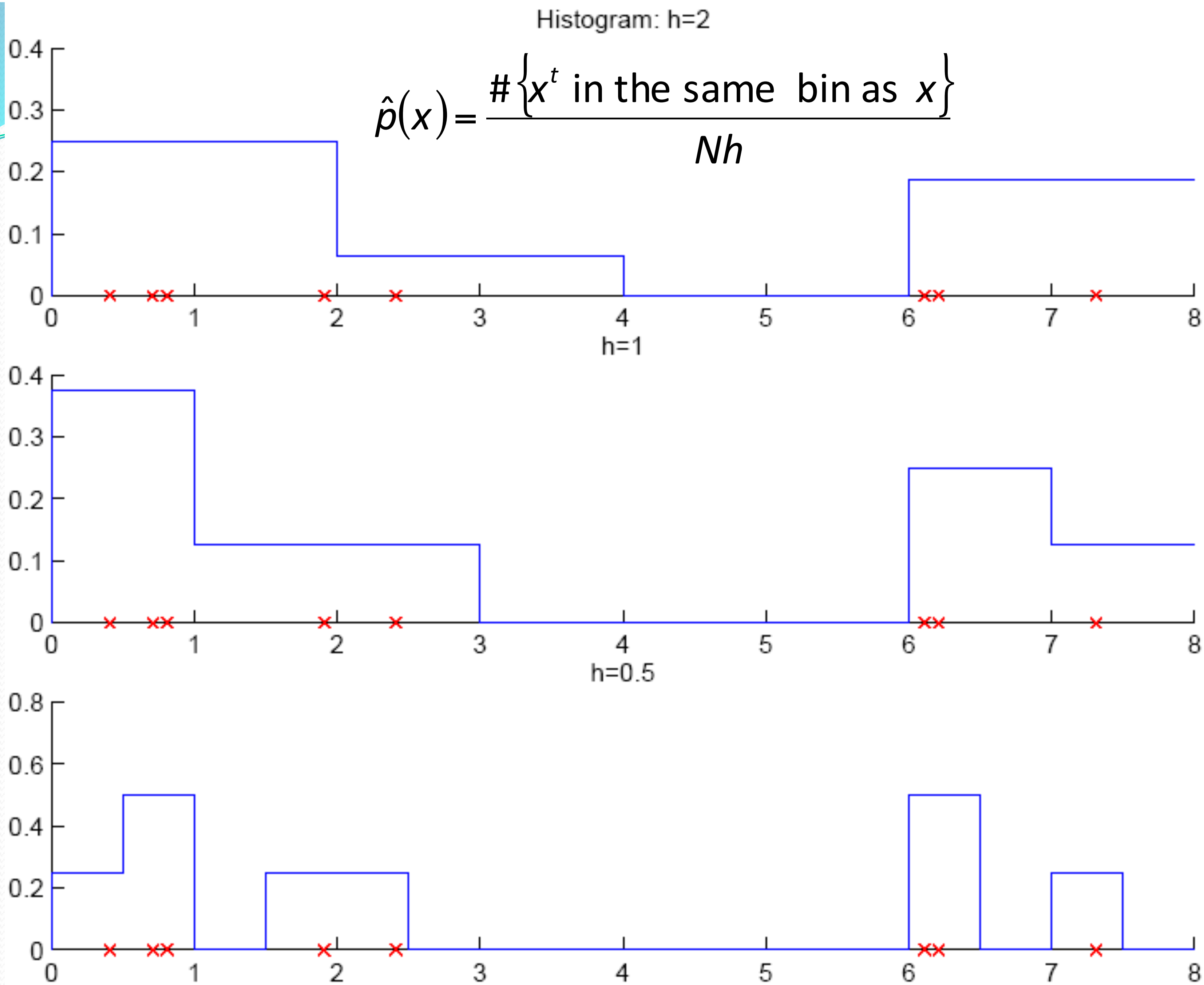
$$\hat{p}(x)=\frac{\#\{x-h<x^t\leq x+h\}}{2Nh}$$

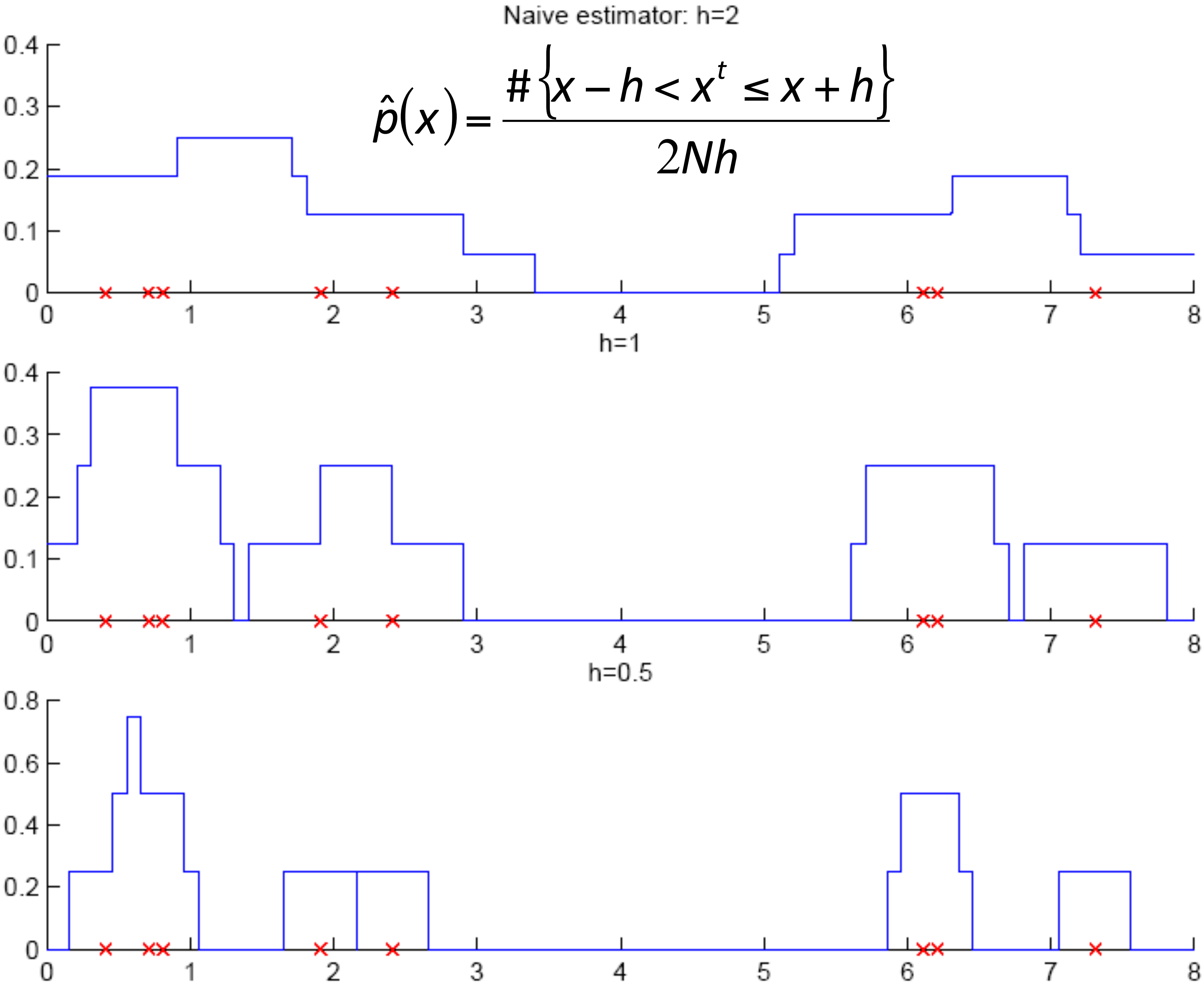
or

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



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





$$\hat{p}(x) = \frac{\#\{x - h < x^t \le x + h\}}{2Nh}$$

Why not histograms

very simple but:

- The final shape of the density estimate depends on the starting position of the bins
- For multivariate data, the final shape of the density is also affected by the orientation of the bins
- The discontinuities of the estimate are not due to the underlying density, they are only an artifact of the chosen bin locations
- These discontinuities make it very difficult, without experience, to grasp the structure of the data
- A much more serious problem is the curse of dimensionality, since the number of bins grows exponentially with the number of dimensions
- In high dimensions we would require a very large number of examples or else most of the bins would be empty

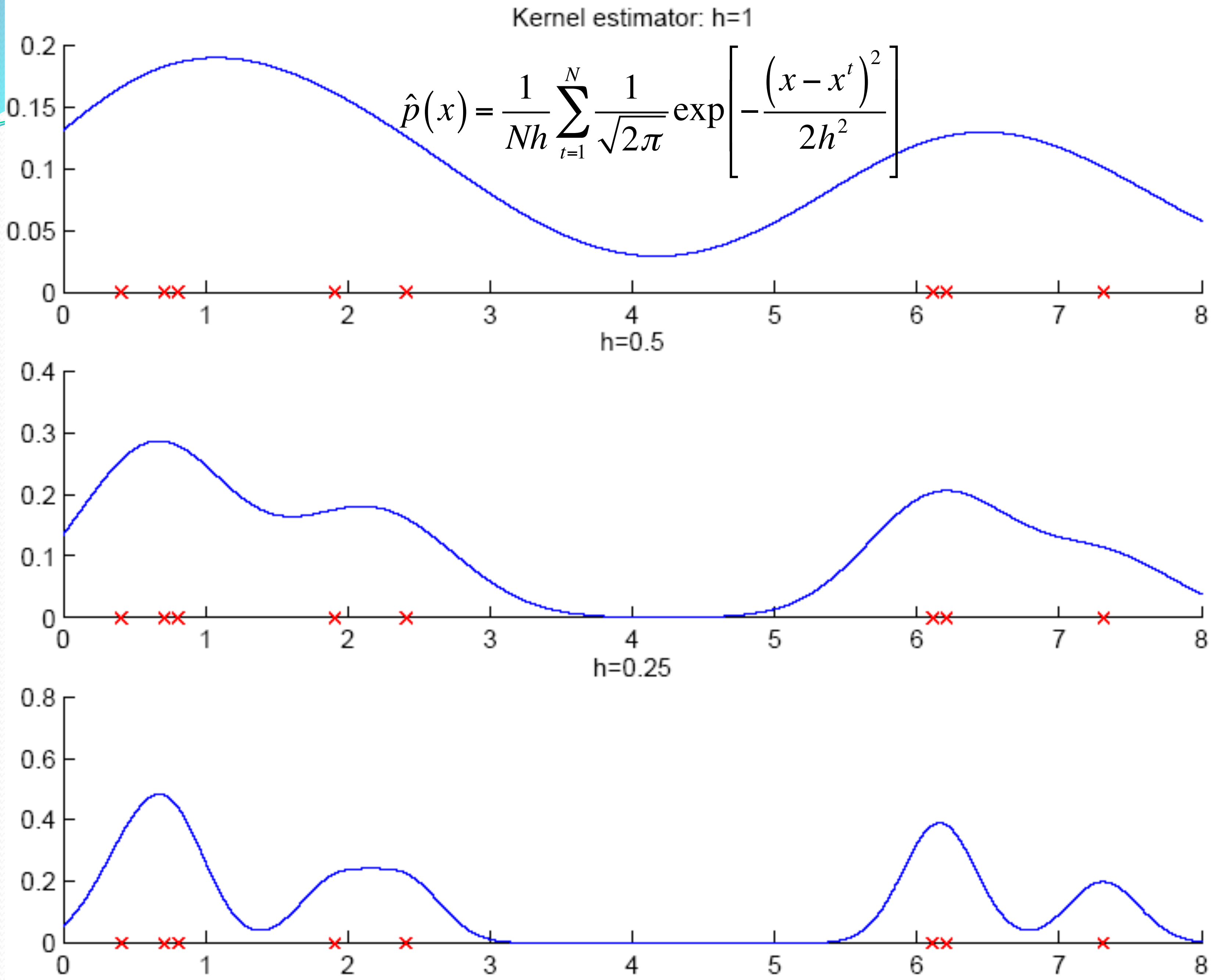
Kernel Estimator

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

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

- Kernel estimator (Parzen windows)

$$\hat{p}(x) = \frac{1}{Nh} \sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right) = \frac{1}{Nh} \sum_{t=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\left(x - x^t\right)^2}{2h^2}\right]$$



$$\hat{p}(x) = \frac{1}{Nh}\sum_{t=1}^{N}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{\left(x-x^t\right)^2}{2h^2}\right]$$

Nonparametric Density Estimation General Formulation (1)

- The probability that a vector x, drawn from a distribution p(x), will fall in a given region $\Re$ of the sample space is

$$P = \int_R p(x')dx'$$

- Suppose now that N vectors {x(1, x(2, …, x(N)} are drawn from the distribution. The probability that k of these N vectors fall in $\Re$ is given by the binomial distribution

$$P(k) = \binom{N}{k} P^k (1-P)^{N-k}$$

- It can be shown (from the properties of the binomial p.m.f.) that the mean and variance of the ratio k/N are
- $$E\left[\frac{k}{N}\right] = P \quad \text{and} \quad Var\left[\frac{k}{N}\right] = E\left[\left(\frac{k}{N} - P\right)^2\right] = \frac{P(1-P)}{N}$$

Nonparametric Density Estimation General Formulation(2)

Therefore, as $N \to \infty$, the distribution becomes sharper (the variance gets smaller) so we can expect that a good estimate of the probability P can be obtained from the mean fraction of the points that fall within $\Re$

$$P \cong \frac{k}{N}$$

- On the other hand, if we assume that $\Re$ is so small that p(x) does not vary appreciably within it, then

$$\int_R p(x')dx' = p(x)V$$

  - where V is the volume enclosed by region $\Re$

Nonparametric Density Estimation General Formulation(3)

Merging with the previous result we obtain

$$\left.\begin{array}{lcl} P = \int_R p(x')dx' & = & p(x)V \\ P & \cong & \dfrac{k}{N} \end{array}\right\} \Rightarrow p(x) \cong \frac{k}{NV}$$

This estimate becomes more accurate as we increase the number of sample points N and shrink the volume V

In practice the value of N (the total number of examples) is fixed

In order to improve the accuracy of the estimate p(x) we could let V approach zero but then the region $\Re$ would then become so small that it would enclose no examples

This means that, in practice, we will have to find a compromise value for the volume V

- Large enough to include enough examples within $\Re$
- Small enough to support the assumption that p(x) is constant within $\Re$

Nonparametric Density Estimation General Formulation(4)

- When applying this result to practical density estimation problems, two basic approaches can be adopted

- We can choose a fixed value of the volume V and determine k from the data. This leads to methods commonly referred to as **Kernel Density Estimation (KDE)**
- We can choose a fixed value of k and determine the corresponding volume V from the data. This gives rise to the **k Nearest Neighbor (kNN)** approach

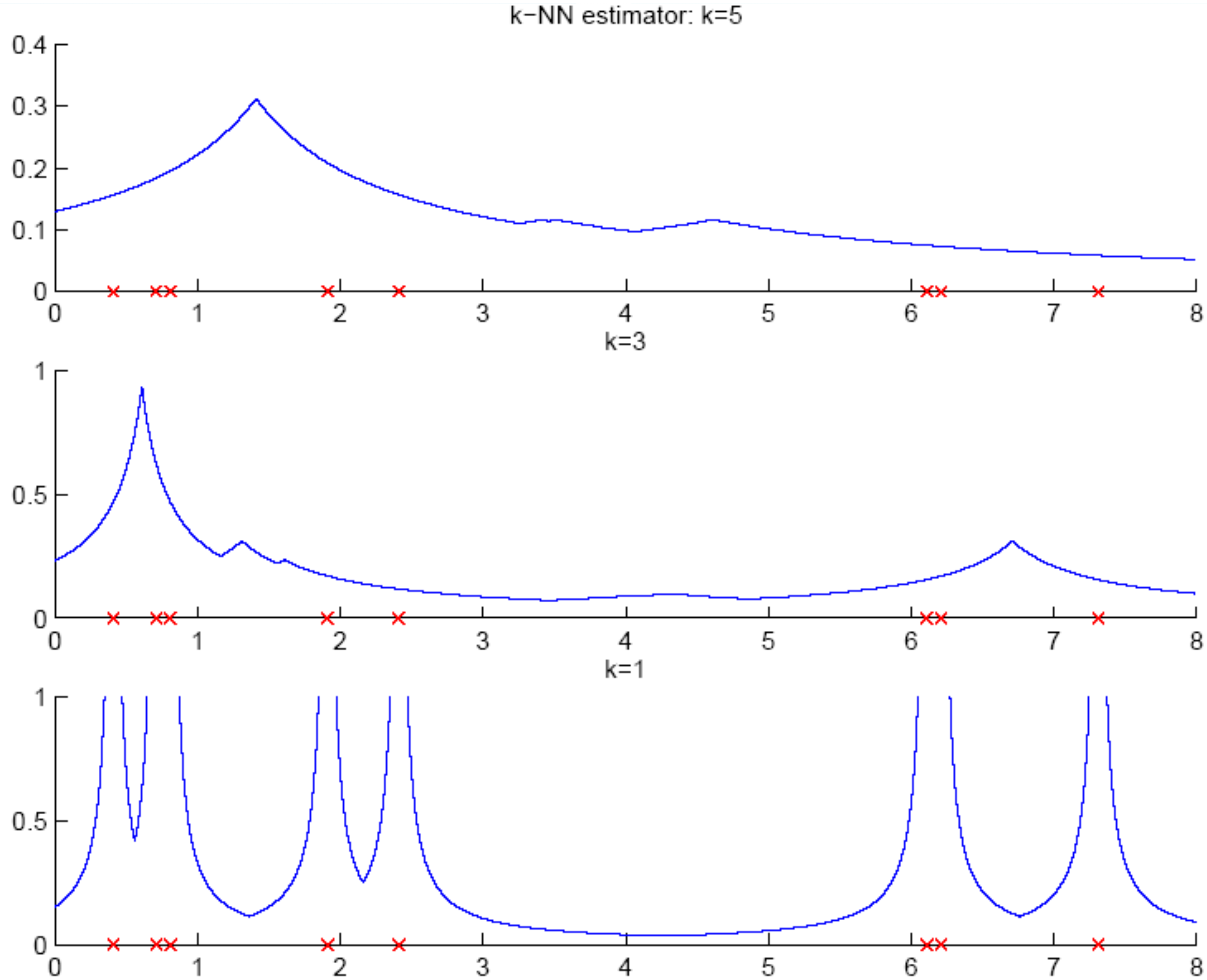
- It can be shown that both kNN and KDE converge to the true probability density as $N \rightarrow \infty$, provided that V shrinks with N, and k grows with N appropriately.

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

$$\hat{p}(x) = \frac{k}{2Nd_k(x)}$$

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


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

Multivariate Data

- Kernel density estimator

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

Multivariate Gaussian kernel

spheric

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

ellipsoid

$$K(\mathbf{u}) = \frac{1}{(2\pi)^{d/2}|\mathbf{S}|^{1/2}} \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{S}^{-1}\mathbf{u}\right]$$

Nonparametric Classification

- Estimate $p(\boldsymbol{x}|C_i)$ and use Bayes' rule
- Kernel estimator

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

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

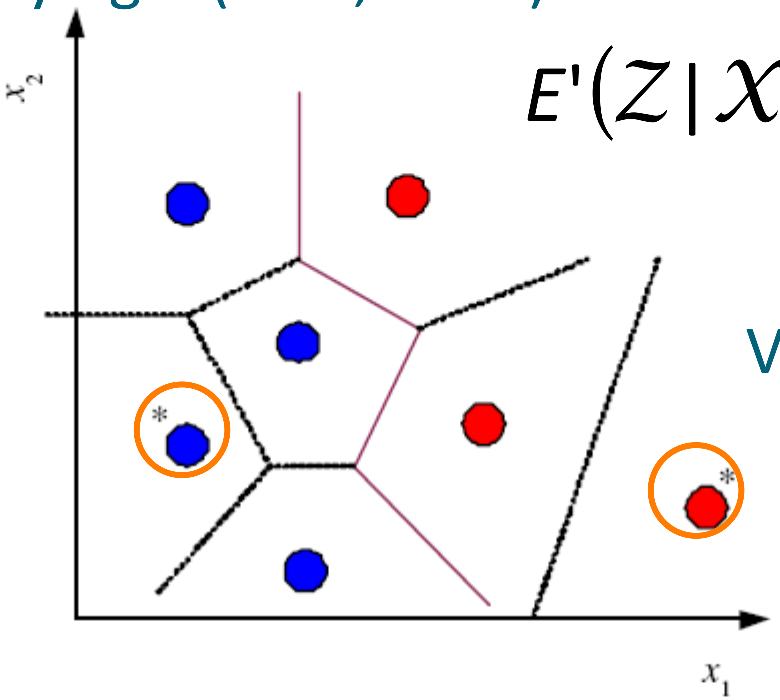
- k-NN estimator

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

Condensed Nearest Neighbor

- Time/space complexity of *k*-NN is O (*N*)
- Find a subset Z of X that is small and is accurate in classifying X (Hart, 1968)

$$E'(\mathcal{Z} \mid \mathcal{X}) = E(\mathcal{X} \mid \mathcal{Z}) + \lambda |\mathcal{Z}|$$



Voronoi Tessalation

Condensed Nearest Neighbor

- Incremental algorithm: Add instance if needed

$\mathcal{Z} \leftarrow \emptyset$

Repeat

For all $\boldsymbol{x} \in \mathcal{X}$ (in random order)

Find $\boldsymbol{x}' \in \mathcal{Z}$ s.t. $\|\boldsymbol{x} - \boldsymbol{x}'\| = \min_{\boldsymbol{x}^j \in \mathcal{Z}} \|\boldsymbol{x} - \boldsymbol{x}^j\|$

If class($\boldsymbol{x}$)$\neq$class($\boldsymbol{x}'$) add $\boldsymbol{x}$ to $\mathcal{Z}$

Until $\mathcal{Z}$ does not change

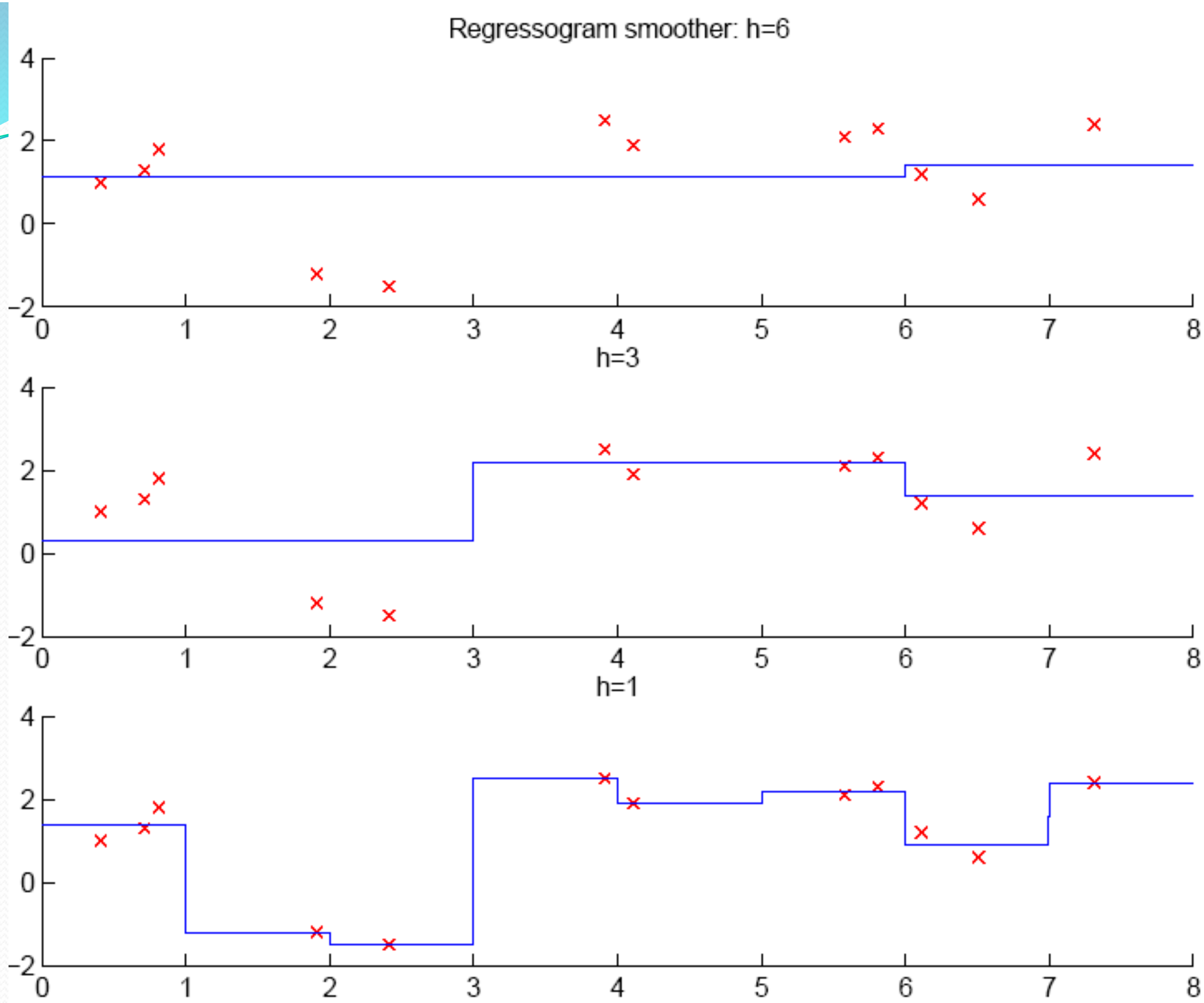
Nonparametric Regression

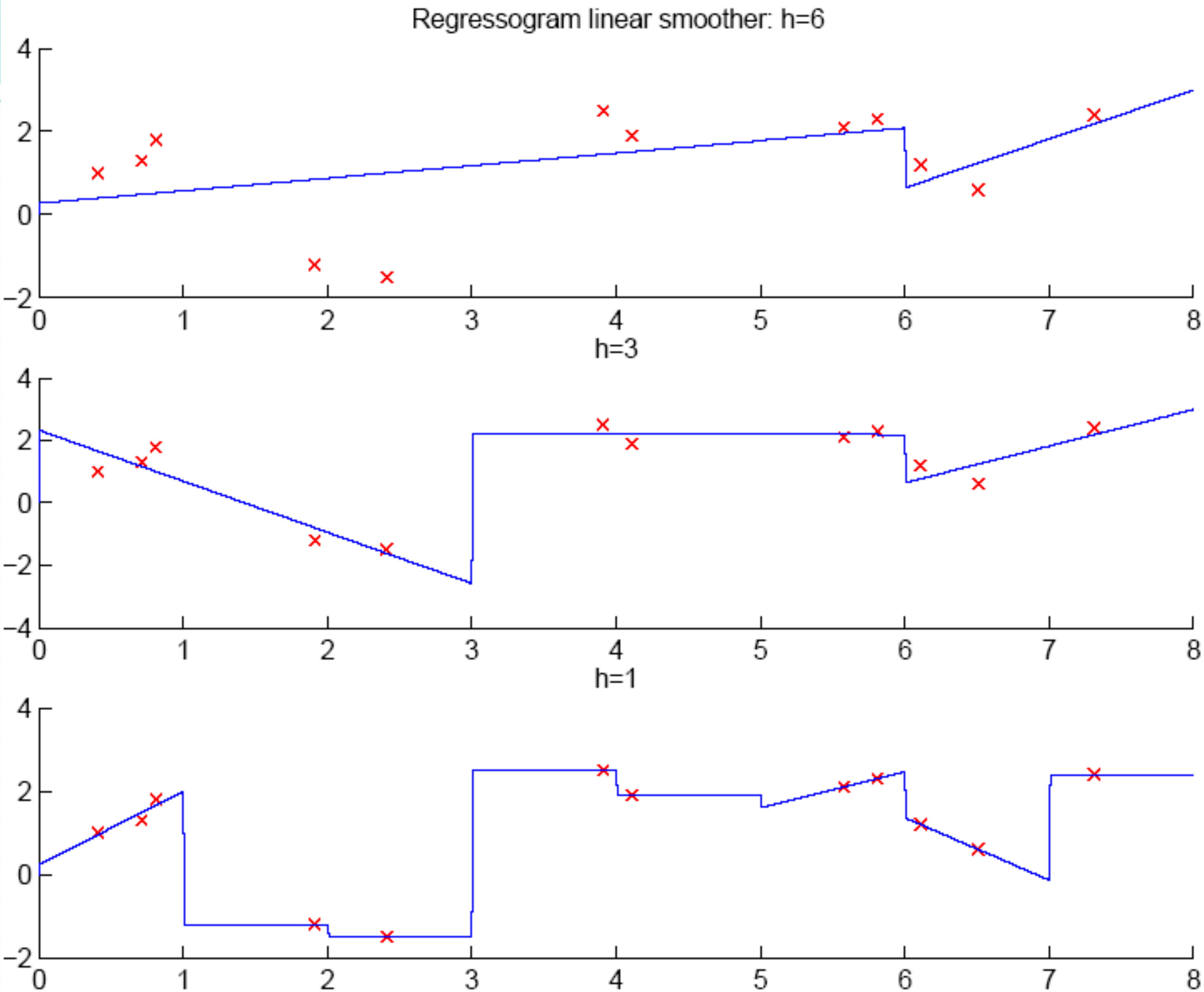
- Aka smoothing models
- Regressogram

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} b(x, x^t) r^t}{\sum_{t=1}^{N} b(x, x^t)}$$

where

$$b(x, x^t) = \begin{cases} 1 & \text{if } x^t \text{ is in the same bin with } x \\ 0 & \text{otherwise} \end{cases}$$


Regressogram smoother: h=6
h=3
h=1
4
2
0
−2
0 1 2 3 4 5 6 7 8


Regressogram linear smoother: h=6
h=3
h=1
4
2
0
−2
−4
0 1 2 3 4 5 6 7 8

Running Mean/Kernel Smoother

- Running mean smoother

$$\hat{g}(x) = \frac{\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right) r^t}{\sum_{t=1}^{N} w\left(\frac{x - x^t}{h}\right)}$$

where

$$w(u) = \begin{cases} 1 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

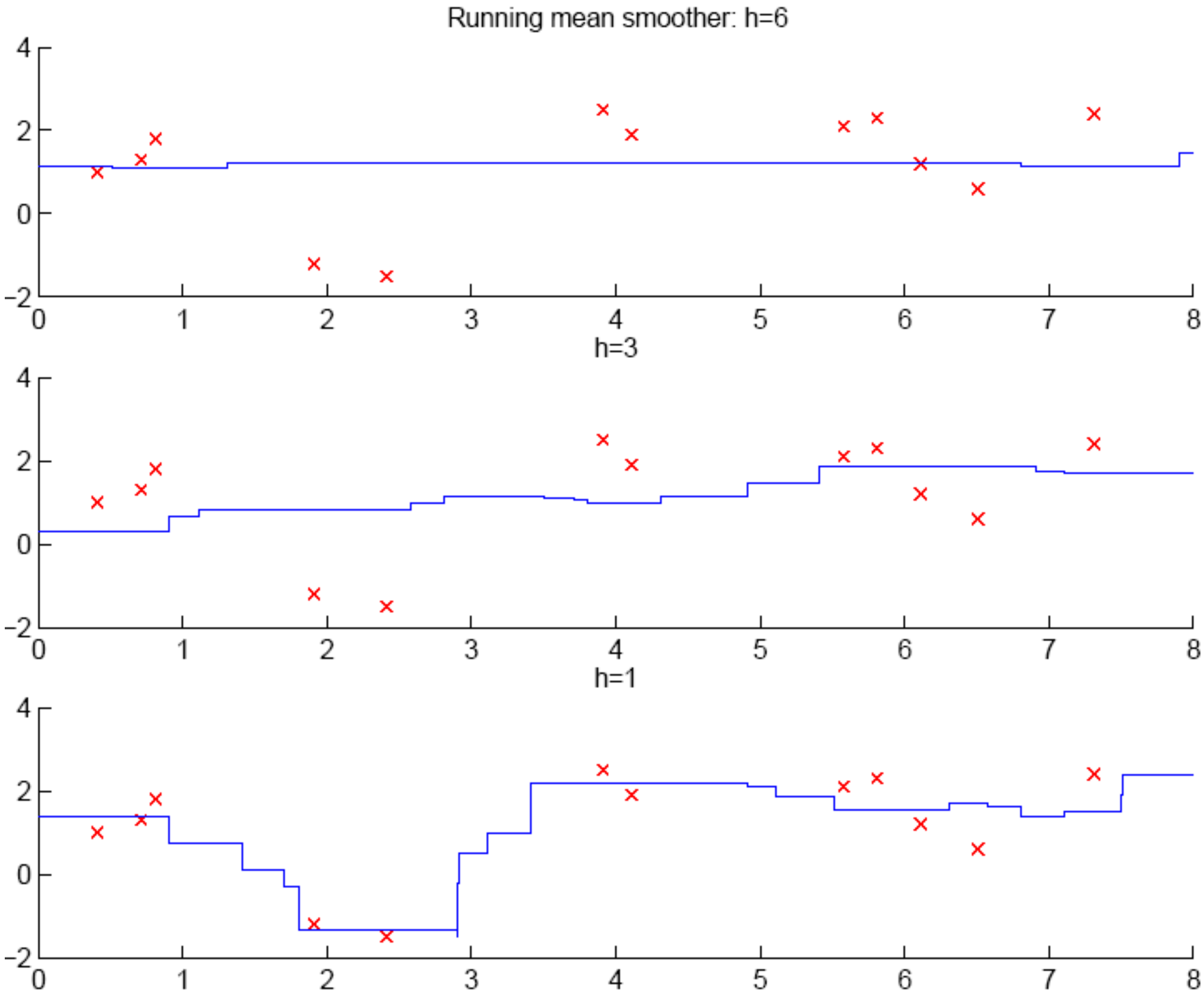
- Running line smoother

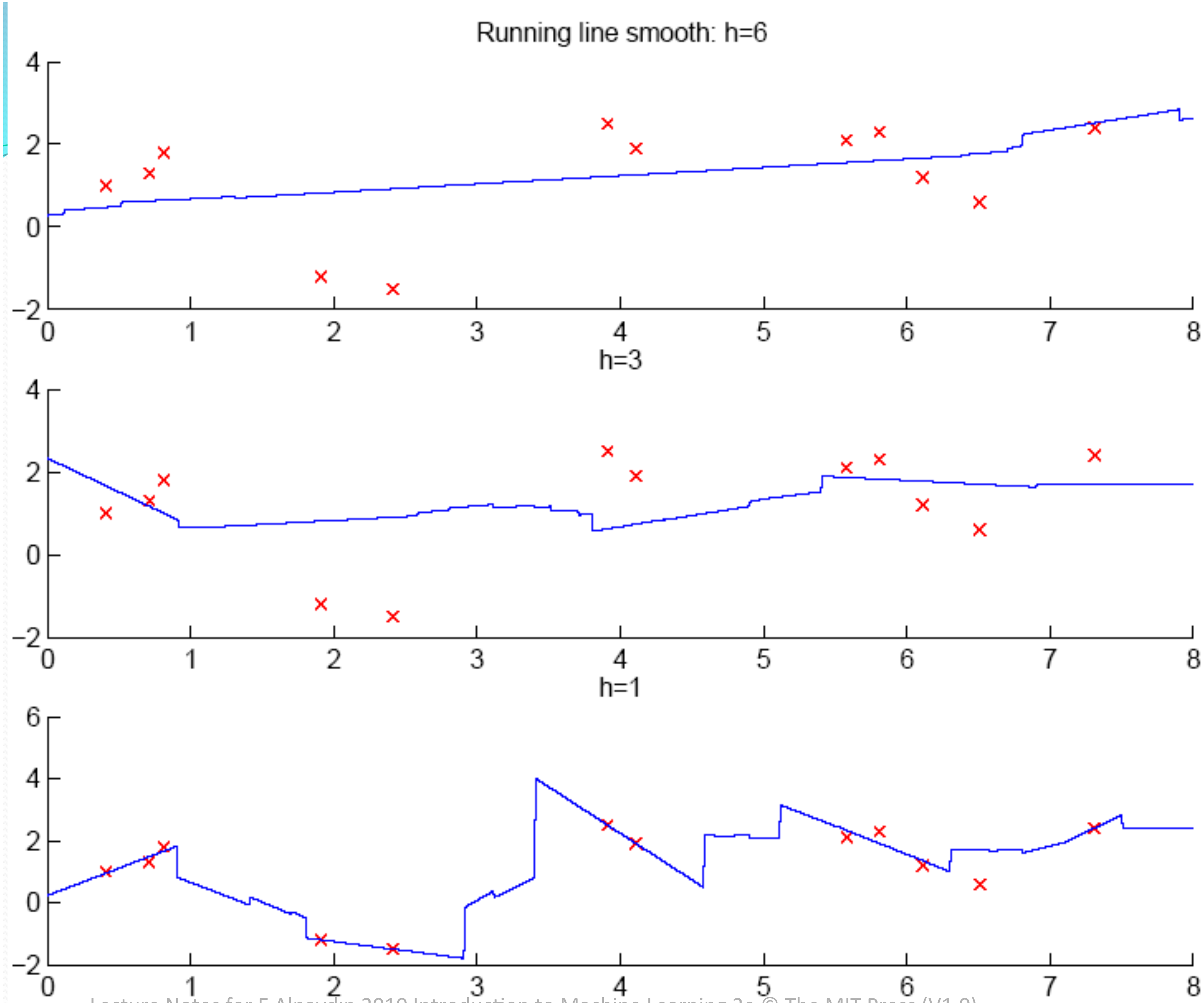
- Kernel smoother

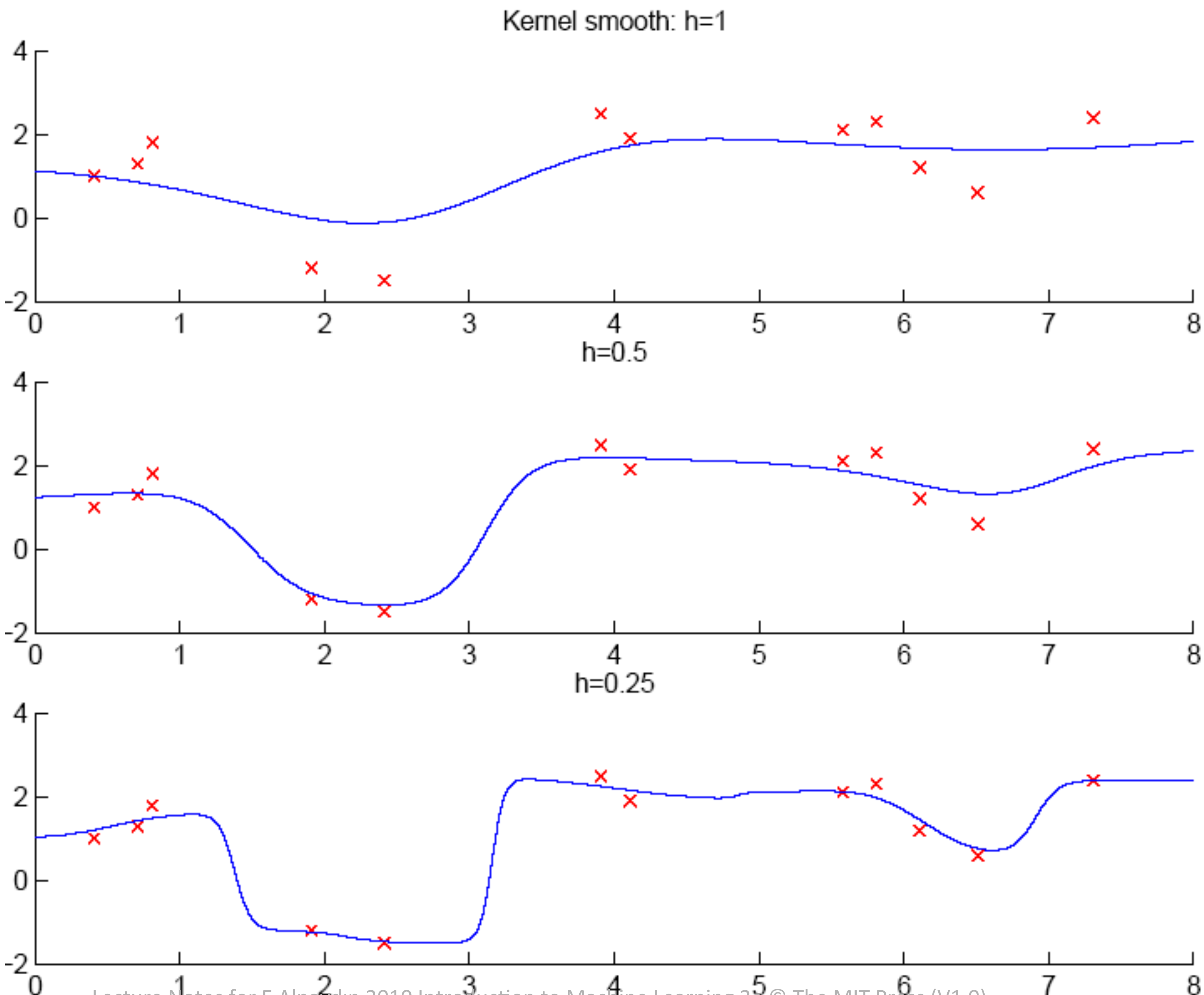
$$\hat{g}(x) = \frac{\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right) r^t}{\sum_{t=1}^{N} K\left(\frac{x - x^t}{h}\right)}$$

where $K(\ )$ is Gaussian

- Additive models (Hastie and Tibshirani, 1990)

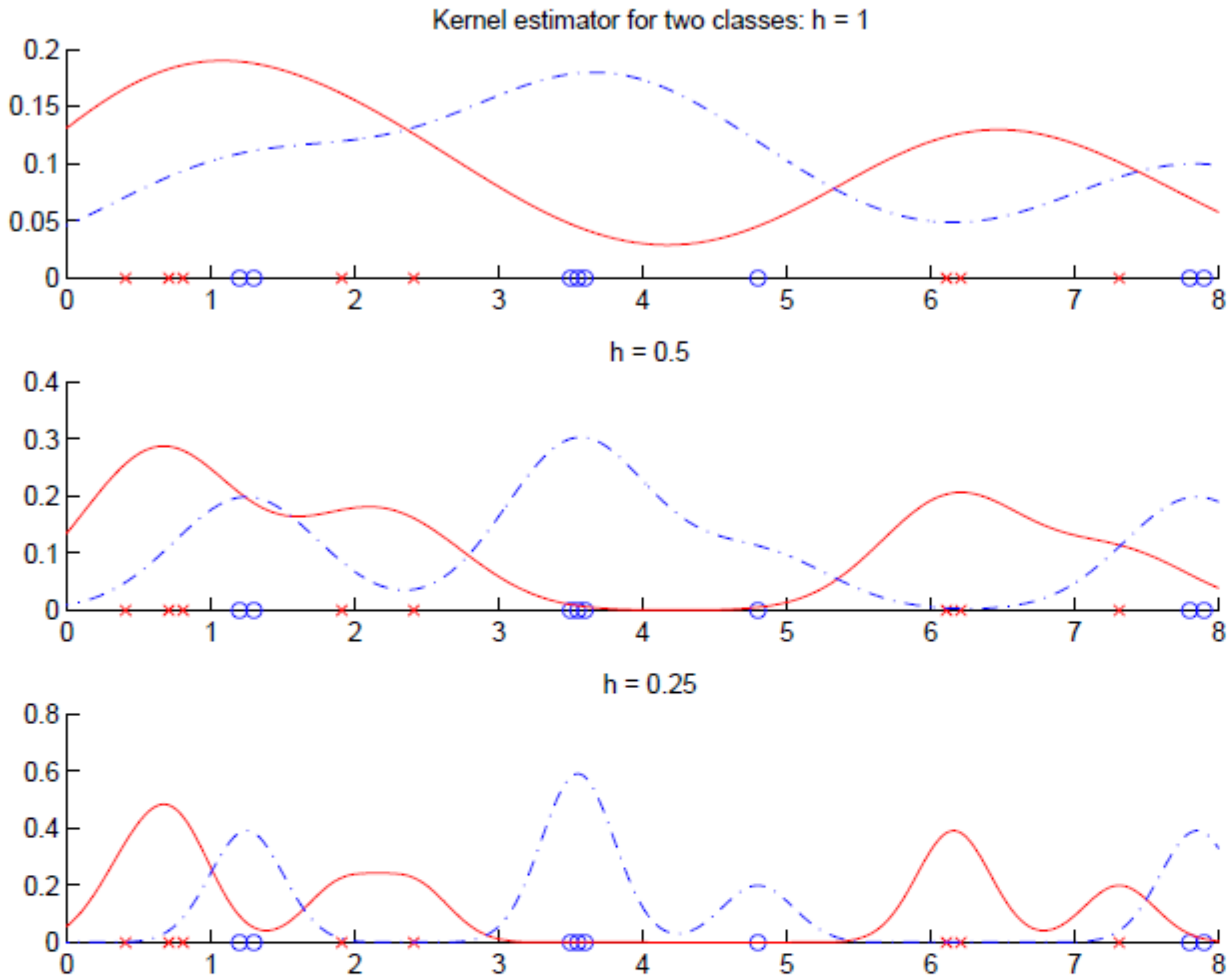

Running mean smoother: h=6
h=3
h=1
4
2
0
-2
0
1
2
3
4
5
6
7
8


Running line smooth: h=6
h=3
h=1
4
2
0
−2
6
0
1
2
3
4
5
6
7
8


Kernel smooth: h=1
h=0.5
h=0.25
4
2
0
-2
0
1
2
3
4
5
6
7
8

How to Choose k or h?

- When *k* or *h* is small, single instances matter; bias is small, variance is large (undersmoothing): High complexity
- As *k* or *h* increases, we average over more instances and variance decreases but bias increases (oversmoothing): Low complexity
- Cross-validation is used to finetune *k* or *h*.


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

Nonparametric Methods for Classification/Regression?

Mostly used models:

- Classification : knn
- Regression: Parzen windows