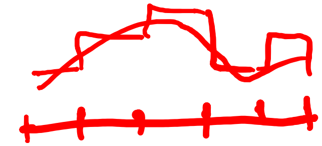
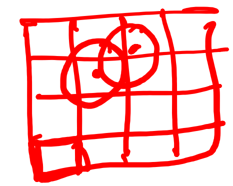
Nonparametric density estimation





k

- Histogram $$\widehat{p}(x) = \frac{n_i}{n\Delta}\mathbf{1}_{x\in\text{Bin}_i}$$

$\Delta_{k,x}$

- Kernel density est $$\widehat{p}(x) = \frac{n_x}{n\Delta}$$

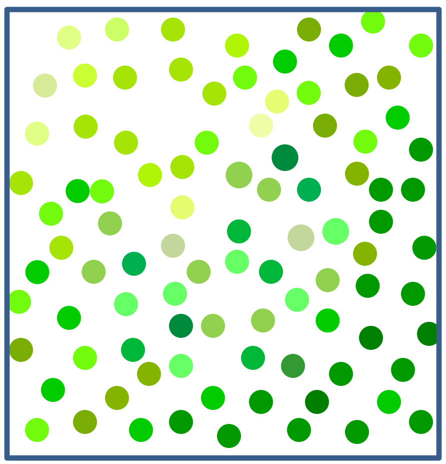
Fix Δ, estimate number of points within Δ of x (n_i or n_x) from data

Fix $n_x = k$, estimate Δ from data (volume of ball around x that contains k training pts)

- k-NN density est $$\widehat{p}(x) = \frac{k}{n\Delta_{k,x}}$$

Local Kernel Regression

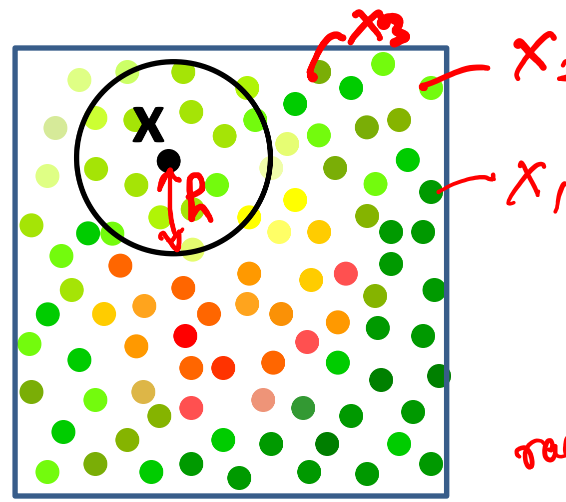
- What is the temperature in the room?



$$\widehat{T} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Global **Average**

at location x?



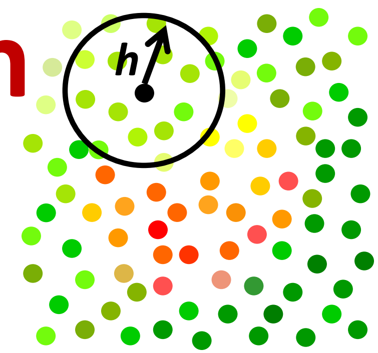
radius vol

$h \equiv \Delta$

$$\widehat{T}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}_{||X_i - x|| \le h}}{\sum_{i=1}^{n} \mathbf{1}_{||X_i - x|| \le h}}$$

"Local" Average

Local Kernel Regression



- Nonparametric estimator
- Nadaraya-Watson Kernel Estimator = local average

$$\sum_{i=1}^{n} w_i(X) = 1$$

h h

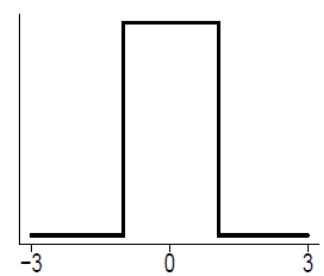
$$\widehat{f}_n(X) = \underbrace{\sum_{i=1}^{n} w_i Y_i}_{\text{local average}} \quad \text{Where} \quad w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)}$$

- Weight each training point based on distance to test point
- Boxcar kernel yields

  local average

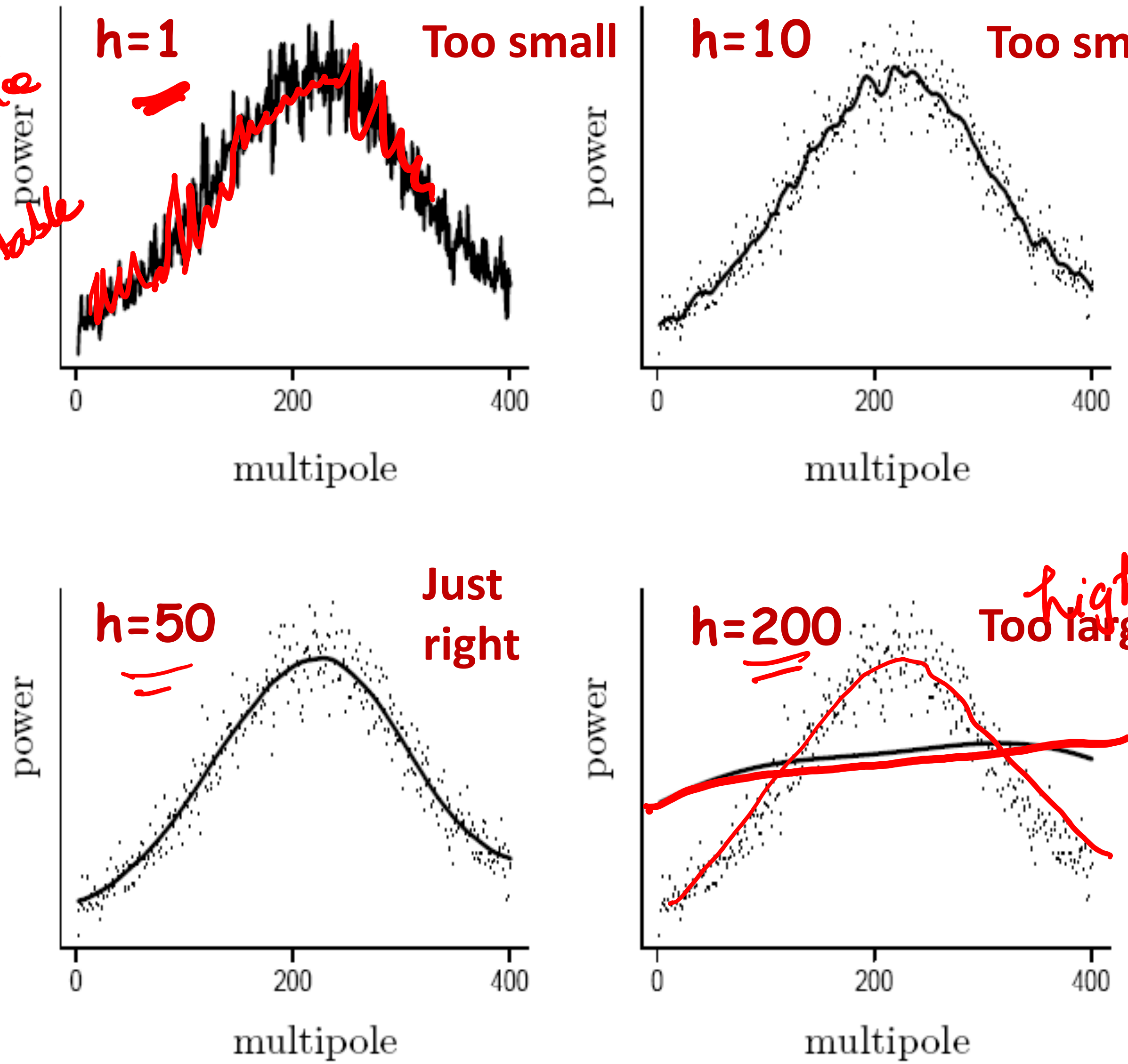
boxcar kernel :

$$K(x) = \frac{1}{2} I(x),$$



Choice of kernel bandwidth h

higher variance ≡ less stable



higher bias ≡ poor approximation

Image Source: Larry's book – All of Nonparametric Statistics local

Kernel Regression as Weighted Least Squares

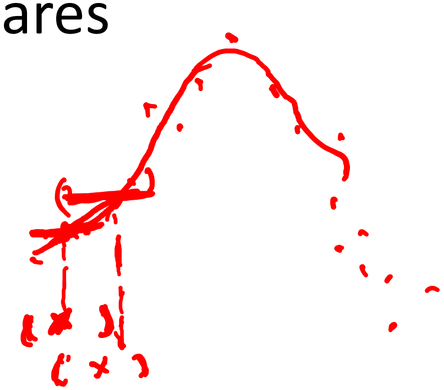
$$\min_{f} \sum_{i=1}^{n} w_i (f(X_i) - Y_i)^2 \qquad\qquad w_i(X) = \frac{K\left(\frac{X-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)}$$

$w_i(X)$

Weighted Least Squares

Kernel regression corresponds to locally constant estimator obtained from (locally) weighted least squares

i.e. set $f(X_i) = \beta$ (a constant)



Kernel Regression as Weighted Least Squares

$K \geq 0$

$\int K = 1$

set $\underline{f(X_i) = \beta}$ (a constant)

$$\hat{\beta}_X \leftarrow \underbrace{\min_{\beta} \sum_{i=1}^{n} \underset{w_i(X)}{w_i}(\overset{f(X_i)}{\overset{\|}{\beta}} - Y_i)^2}_{}$$

constant

$$\sum_{i=1}^{n} w_i(X) = \frac{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)} = 1$$

$$\frac{\partial J(\beta)}{\partial \beta} = 2\sum_{i=1}^{n} w_i(\beta - Y_i) = 0$$

$$\beta \underbrace{\sum_{i=1}^{n} w_i}_{1} = \boxed{\sum_{i=1}^{n} w_i Y_i}$$

$$\Rightarrow \widehat{f}_n(X) = \widehat{\beta}_X = \sum_{i=1}^{n} w_i Y_i$$

Notice that $\sum_{i=1}^{n} w_i = 1$

Local Linear/Polynomial Regression

$X_i\beta$

$$\min_f \sum_{i=1}^{n} w_i(f(X_i) - Y_i)^2 \qquad w_i(X) = \frac{K\left(\frac{X-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)}$$

Weighted Least Squares

Local Polynomial regression corresponds to locally polynomial estimator obtained from (locally) weighted least squares

i.e. set $f(X_i) = \beta_0 + \beta_1(X_i - X) + \frac{\beta_2}{2!}(X_i - X)^2 + \cdots + \frac{\beta_p}{p!}(X_i - X)^p$

(local polynomial of degree p around X)

Summary

- Non-parametric approaches

Four things make a nonparametric/memory/instance based/lazy learner:

1. *A distance metric, dist(x,X_i)*

   Euclidean (and many more)

   $K\left(\frac{\|x - x_i\|}{h}\right)$ $K\left(\frac{d(x, x_i)}{h}\right)$

2. *How many nearby neighbors/radius to look at?*

   k, Δ/h

3. *A weighting function (optional)*

   W based on kernel K

   $W = \frac{K}{\sum K}$

4. *How to fit with the local points?*

   Average, Majority vote, Weighted average, Poly fit

Summary

- Parametric vs Nonparametric approaches
  - Nonparametric models place very mild assumptions on the data distribution and provide good models for complex data

    Parametric models rely on very strong (simplistic) distributional assumptions

  - Nonparametric models (not histograms) requires storing and computing with the entire data set.

    Parametric models, once fitted, are much more efficient in terms of storage and computation.