

# Lecture 14

03/03/2022

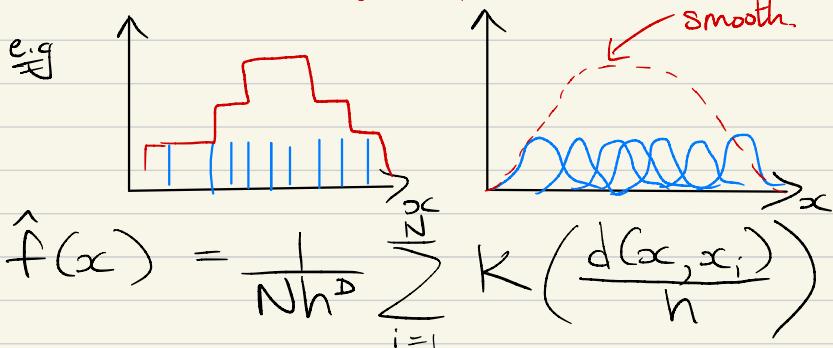
## Density Estimation

### \* Non-parametric density estimation

- ⇒ we know nothing about the data.
- ⇒ **no** functional form is fit to the distribution of events.

### (1) Kernel Density Estimation

↳ blur and smooth by replacing each point/sample by a distribution or "KERNEL".



$d$  = distance metric (e.g.  $(x - x_i)$ )

$h$  = bandwidth

$D$  = dimension of data

$N$  = number of samples.

### Choice of kernels

GAUSSIAN

$$K(x) = \frac{1}{(2\pi)^{\frac{D}{2}}} e^{-\frac{1}{2} \frac{d^2}{h^2}}$$

### EPANECHNIKOV

$$K(x) = \frac{3}{4} \left(1 - \frac{d^2}{h^2}\right)$$

How do we choose bandwidth?

↳ cross-validation

PARTITION  
DATA

{ → training → fit KDE model  
→ validation → tune bandwidth  
→ testing → final performance check

## (2) Nearest Neighbor Density Estimation

↳ estimate density by comparing distances to set of nearest neighbouring samples.

$$\hat{f}_K(x) = \frac{c}{\sum_{i=1}^K d_i}$$

constant evaluated at end through normalization.

## \* Parametric Density Estimation

⇒ fit a functional form to the density of samples.

e.g. Gaussian Mixture Models.

→ KDEs replace each point with a kernel function!

→ Instead, GMMs model the distribution with a **SUM OF GAUSSIANS** at different locations, and with different widths

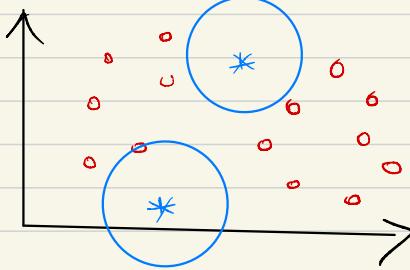
$$\text{i.e. } \hat{f}_{\text{GMM}}(x) = \sum_{k=1}^N (\alpha_k N(x | \mu_k, \Sigma_k))$$

all these have  
to be estimated

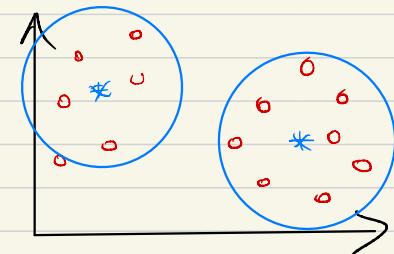
- Clustering  $\Rightarrow$  unsupervised, since we don't have labels.

### \* K-Means

- choose a number of clusters.
- choose random initial cluster centers
- decide on which cluster each sample belongs to
- $$C_k(x_i) = \arg \min_k \|x_i - \mu_k\|$$
- minimize 
$$\sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|^2$$
- ITERATE



INITIAL



BEST-FIT

## \* Mean-shift Clustering

- compute KDE of sample set
- move all points in the gradient direction.
- iterate until all points have reached maxima.

⇒ don't need to decide on # clusters.

## \* Correlation Functions

⇒ On what scales does a distribution of samples differ from a completely random, uncorrelated sample.

- e.g.\* angular scale of CMB temperature fluctuations
- \* timescales of signal or noise correlations in time-domain data.

## DENSITY FLUCTUATION OF SOURCE COUNTS (e.g. galaxies)

$$\hookrightarrow \xi(r) = \left\langle \frac{S_p(x)}{P} \cdot \frac{S_p(x+r)}{P} \right\rangle$$

-- why 2-point correlations?

↳ Gaussian fluctuations are ENTIRELY characterized by mean (1-point) and variance (2-point)