Lecture 14

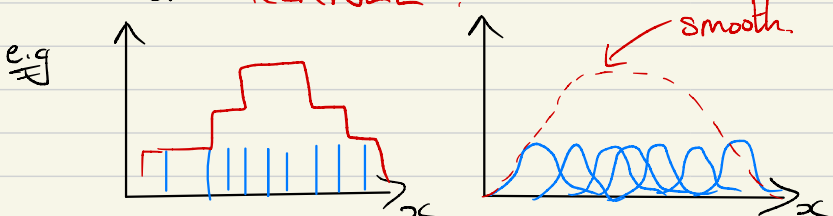
Density Estimation

* 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".



$$\hat{f}(x) = \frac{1}{Nh^D} \sum_{i=1}^{N} K\left(\frac{d(x, x_i)}{h}\right)$$

d = distance metric (e.g $(x - x_i)$)
h = bandwidth
D = dimension of data
N = number of samples.

Choice of Kernel

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

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 neighboring samples.

$$\hat{f}_k(x) = \frac{C}{\sum_{i=1}^{k} d_i^D}$$

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

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

all these have to be estimated

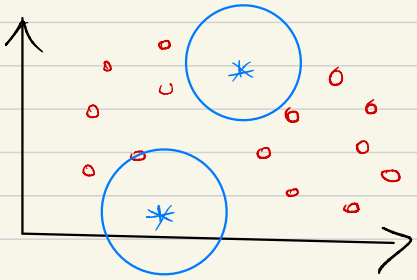
- Clustering ⇒ unsupervised, since we don't have labels.

* K-Means

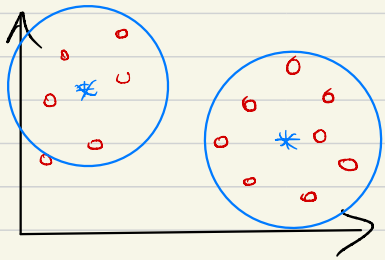
- choose a number of clusters.
- choose random initial cluster centers
- decide on which cluster each sample belongs to ...

$$C_k(x_i) = \arg\min_k \| x_i - \mu_k \|$$

- 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 +ve 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 (eg. galaxies)

↳ $\xi(r) = \left\langle \frac{\delta\rho(x)}{\rho} \cdot \frac{\delta\rho(x+r)}{\rho} \right\rangle$

--- why 2-point correlations?

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