- So far we have seen parametric densities like Gaussian, GMM, etc. which makes an assumption about the form.
- non-parametric estimation – estimate $p(x)$ w/o strong assumptions, using the data. (Note, also has parameters)

Histogram

- Assume samples $\{x_1, \dots, x_n\}$
- Consider a region R
- Define $P = p(x \in R) = \int_{x \in R} p(x)\, dx$

  prob. a point in R
- Define $K_R =$ # points inside R
- Eestimate of P: $\hat{P} = \frac{K_R}{n}$
- Assume R is small, then

  $\hat{P} \approx p(x) \cdot V_R$, $V_R =$ volumn of R,

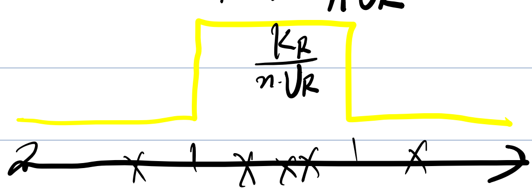
  $x =$ center of R.

  approximate integral over R with rectangle.

$$\hat{P} = \frac{K_R}{n} \qquad \hat{P} = p(x) \cdot V_R$$

$$\rightarrow p(x) \cdot V_R = \frac{K_R}{n}$$

$$p(x) = \frac{K_R}{n \cdot V_R}$$



This is just a histogram, but we can extend it

Q: How to choose R

✓ 1) keep V_R fixed, & let K_R vary → Parzen windows, kernal density estimation

2) keep K_R fixed, & let V_R vary → k-NN estimation (k nearest neighbors)

in general 1) is better, why?

Second one requires the same number of points in each region. for a low density region (region should be extremely large)

dense region (region will be too small)

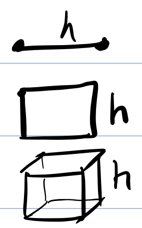
Kernal Density Estimation.

- let R be a d-dim hypercube w/ side of h

d=1

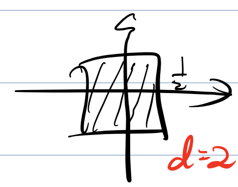
d=2

d=3



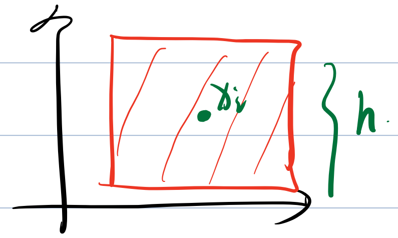
- introduce a window

$$K(x) = \begin{cases} 1, & |x_i| \le \frac{1}{2}, \ \forall i = \{1, \dots, d\} \\ 0, & \text{otherwise} \end{cases}$$

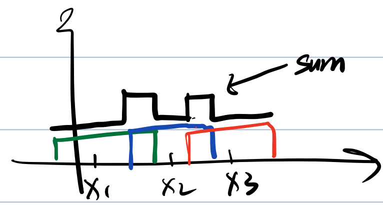


(Parzen window → kernal function)

$$K\left(\frac{x - x_i}{h}\right) = \begin{cases} 1, & \text{if } x \text{ falls inside a cube w/ side } h, \text{ centered at } x_i \\ 0, & \text{otherwise} \end{cases}$$



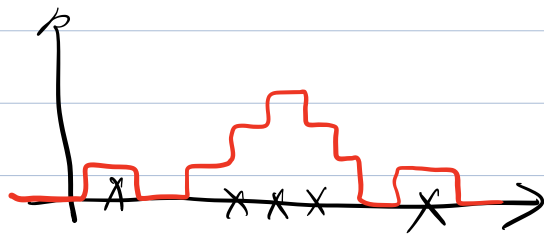
– # of points near x: $k = \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$



"Stacking boxes centered at all x_i's"

$$\hat{p}(x) = \frac{1}{n}\frac{k_R}{V_R}$$

$$\hat{p}(x) = \frac{1}{n \cdot h^d}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$



Other kernel function.

Constraints: $\left.\begin{array}{l} K(x) \geq 0 \\ \int K(x)\,dx = 1 \end{array}\right\}$ i.e. valid pdf

Examples:

Uniform $K(x) = \begin{cases} 1, & |x_i| \leq \frac{1}{2} \ \forall i \\ 0, & \text{otherwise.} \end{cases}$

Unit sphere $K(x) = \begin{cases} \frac{1}{c}, & \|x\|^2 \leq 1 \\ 0, & 0 \end{cases}$ – c is the volumn

Gaussian: $k(x) = \frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}\|x\|^2}$ (assume $\sigma^2 = 1$)

$\hat{p}(x) = \frac{1}{n \cdot h^d} \cdot \sum_i k\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_i \mathcal{N}(x \mid x_i, h^2 I)$

π_i — mean, ith comp.

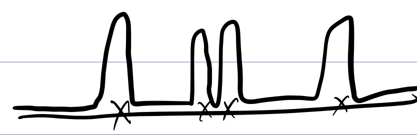
GMM w. n components

Bandwidth parameter h.

h controls the smoothness of $\hat{p}$

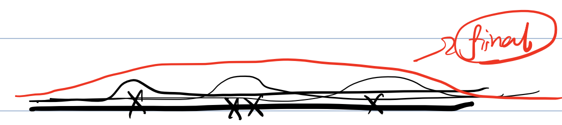
Intuitively

h too small



noisy estimate if not enough samples

h too large



blurry estimate if too many samples

Convergence Analysis

Will $\hat{p}(x)$ converge to true $p(x)$?

$\hat{p}(x)$ depends on samples $\{x_i\}_i$, which are r.v.s,

$\Rightarrow$ we can look at bias / variance

$\hat{p}(x)$ converges to $p(x)$ if

① $\lim_{n \to \infty} E[\hat{p}(x)] = p(x)$

② $\lim_{n \to \infty} \mathrm{Var}[\hat{p}(x)] = 0$

Define $\tilde{k}(x) = \underbrace{\frac{1}{h^d}}_{\text{scale the amplitude}} \underbrace{k\left(\frac{x}{h}\right)}_{\text{scale width of kernal}} \Rightarrow \hat{p}(x) = \frac{1}{n} \sum_i \tilde{k}(x - x_i)$

Mean: $E[\hat{p}(x)] = E_{x_i}\left[\frac{1}{n} \sum_i \tilde{k}(x - x_i)\right]$

$\vdots$ (tutorial)

$$= \int p(u)\, \tilde{k}(x-u)\, du$$

$$= p(x) * \tilde{k}(x)$$

convolution of the true $p(x)$ with the kernal $\tilde{k}(x)$

$\Rightarrow$ blurred version of $p(x)$

Only unbiased when

$$\tilde{k}(x) = \delta(x) = \lim_{h \to 0} \tilde{k}(x - x_i) \Rightarrow E[\hat{p}(x)] = p(x)$$

Dirac delta

$$\int f(x) \cdot \delta(x - x_0)\, dx = f(x_0)$$

(formal definition of δ)

Variance: $Var(\hat{p}(x)) \leq \frac{1}{nh^d} \max_x [k(x)] \cdot E[\hat{p}(x)]$ (tutorial)

For small variance, we need n large / h large.

★ h controls the bias and variance

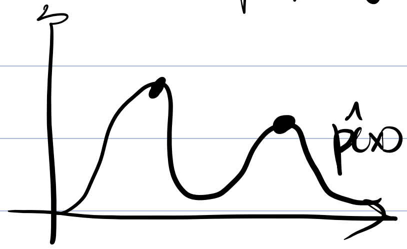
$$\begin{cases} h \to 0 & \text{bias} \to 0, \text{ variance large} \\ h \to \infty & \text{bias} \neq 0, \text{ variance small} \end{cases}$$

How to select h? cross-validation

- select h to maximize LL of validation set
- select h as function of physical property (inherent property of noise)

Mean-Shift algorithm

Find the modes (peaks) of $\hat{p}(x)$

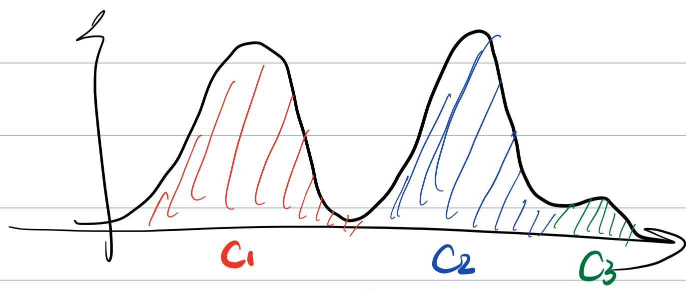


1) Start at a point $\hat{x}$ (e.g. one datapoint x_i)
2) use gradient ascent to move uphill [$\hat{x} \leftarrow \hat{x} + \lambda \nabla \hat{p}(x)$]
3) eventually $\hat{x}$ will converge to a mode.

- Repeat for many different initial $\hat{x}$s to find the modes

Clustering:

The x_i that converge to the same mode belong to the same cluster.



辐射的 "basins of attraction"

Consider radially symmetric kernels.

meaning $$\underbrace{k(x)}_{\text{kernel function}} = \underbrace{\alpha}_{\text{const}} \underbrace{\bar{k}(\|x\|^2)}_{\text{kernel profile}}$$

eg. Gaussian $$k(x) = \frac{1}{(2\pi)^{d/2}} \cdot e^{-\frac{1}{2}\|x\|^2}$$

$$\bar{k}(r) = e^{-\frac{1}{2}r}.$$

$$\alpha = (2\pi)^{-d/2}$$

KDE (kernal density estimation)

$$\hat{p}(x) = \frac{\alpha}{n h^d} \sum_i \bar{k}\left(\left\|\frac{x-x_i}{h}\right\|^2\right)$$

gradient: define $\bar{g}(r) = -\bar{k}'(r)$ (Gaussian $\bar{g}(r) = \frac{1}{2} e^{-\frac{1}{2}r}$)

$$\nabla \hat{p}(x) = \left(\frac{\alpha}{n \cdot h^d} \sum_i \bar{k}\left[\left(\frac{x-x_i}{h}\right)^T \left(\frac{x-x_i}{h}\right)\right]\right)'$$

$$= \frac{\alpha}{n \cdot h^{d+2}} \left[\sum_i \bar{g}\left(\left\|\frac{x-x_i}{h}\right\|^2\right)(2x-2x_i)\right]$$

$$= \underbrace{\frac{2\alpha}{n \cdot h^{d+2}} \left(\sum_i \bar{g}\left(\left\|\frac{x-x_i}{h}\right\|^2\right)\right)}_{\propto \text{ KDE using } \bar{g}(r) \; = \hat{g}(x)} \cdot \Bigg(\underbrace{\frac{\sum_i x_i \bar{g}\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_i \bar{g}\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}}_{\text{weighted mean of samples closed to } x} - x\Bigg)$$

const

kernal profile

large only when $x \to x_i$

"mean-shift" vector

diff between weighted mean inside the window and the center of the window

$= m(x)$

Gradient ascent

$$\hat{x}^{(k+1)} = \hat{x}^{(k)} + \lambda \nabla \hat{p}(\hat{x}^{(k)})$$

updated, current, step size $\Rightarrow$ important for convergence.

Use an adaptive step size

$$\lambda = \frac{1}{\hat{g}(x)}$$

$\hat{g}(x)$ is small $\Rightarrow$ large step size (low-density region)

$\hat{g}(x)$ is large $\Rightarrow$ small step size (high-density region)

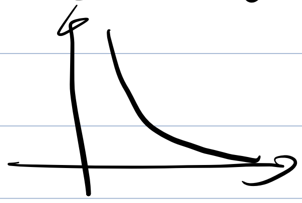
$$\Rightarrow \hat{x}^{(k+1)} = \hat{x}^{(k)} + \frac{1}{\hat{g}(\hat{x}^{(k)})} \hat{g}(\hat{x}^{(k)}) \cdot m(\hat{x}^{(k)})$$ (updated gradient descent)

$$= \hat{x}^{(k)} + m(\hat{x}^{(k)})$$

$$= \frac{\sum x_i \bar{g}\left(\left\|\frac{\hat{x}^{(k)}-x_i}{h}\right\|^2\right)}{\sum \bar{g}\left(\left\|\frac{\hat{x}^{(k)}-x_i}{h}\right\|^2\right)}$$

Intuitively

Note: ① proved in paper, guarenteed to converge to a stationary point if kernal profile $\bar{k}(r)$ is monotonically decreasing & convex



eg. Gaussian