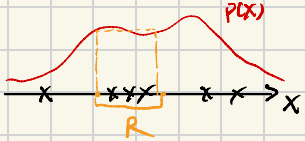
Lecture 5 Non parametric density estimation

So far we have used parametric models - assume a form of the density (Gaussian, Exp, GMM)

Non-parametric estimation - estimate p(x) without assuming a form
Note: non parametric estimation still has parameters

Histogram : samples $\{X_1, \dots X_N\}$
consider region R



Define $\pi_R = p(x \in R) = \int_{x \in R} p(x)dx$
prob : draw sample from R

Define : K_R : # samples in R
MLE : $\hat{\pi}_R = K_R / N$

- Assume R is small (p(x) is flat locally)
$\hat{\pi}_R = p(x) \cdot V_R$ ← volume of R

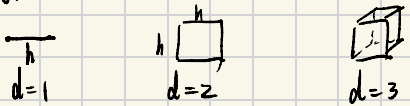
Solve p(x) : $p(x) V_R = \frac{K_R}{N} \Rightarrow \hat{p}(x) \approx \frac{K_R}{N \cdot V_R}$, $x \in R$

How to select R?

1) Keep V_R fixed, let K_R vary — kernel density estimation (KDE)
2) Keep K_R fixed, let V_R vary — k-NN estimator

Kernel Density Estimation

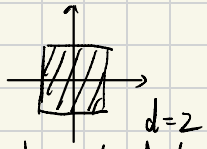
- Hypercube : R d-dim w/ size h



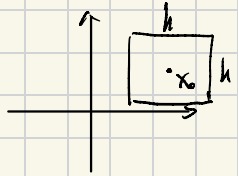
$V_R = h^d$

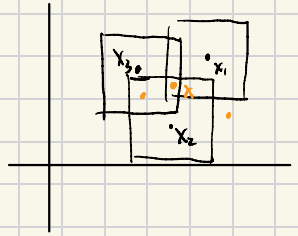
- window function (h = 1)

$$k(x) = \begin{cases} 1, & |x_i| < 1/2, \ \forall i \in \{1 \dots d\} \\ 0, & \text{otherwise} \end{cases}$$



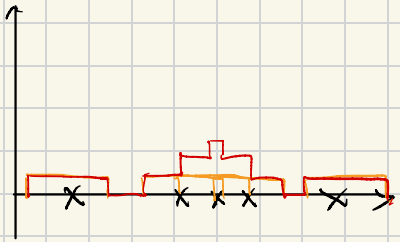
Note: $k\left(\frac{x - x_i}{h}\right) = \begin{cases} 1, & x \text{ is inside hypercube centered at } x_i, \text{ w/ length } h. \\ 0 & \text{otherwise} \end{cases}$





$K = \sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right)$

Hence : $\hat{p}(x) = \frac{1}{N} \frac{k_R}{V_R} = \frac{1}{N h^d} \sum_{i=1}^{N} k\left(\frac{x - x_i}{h}\right)$



other kernel func

$$\left.\begin{array}{l} k(x) \geq 0 \\ \int k(x)\,dx = 1 \end{array}\right\} \Rightarrow \text{pdf}$$

Example:

unix box $k(x) = \begin{cases} 1 & |x_i| \leq \frac{1}{2} \quad \forall i \in \{1,\dots,d\} \\ 0 & \text{Other} \end{cases}$

unit sphere $k(x) = \begin{cases} \frac{1}{\alpha} & \|x\|^2 < 1 \\ 0 & \end{cases}$

α: volumn of Hypersphere

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

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

GMM π_j mean μ_j cov Σ_j

bandwidth param

h controls the smoothness of $\hat{p}(x)$

h too small

Noisy estimate if no enough points

h too large

Blurred estimate if the points are too close

Convergence Analysis

- Will $\hat{p}(x)$ converge to the $p(x)$?
- $\hat{p}(x)$ converges to $p(x)$ if 1) $\lim_{N\to\infty} \mathbb{E}[\hat{p}(x)] = p(x)$

  2) $\lim_{N\to\infty} \text{var}[\hat{p}(x)] = 0$

Define: $\tilde{K}(x) = \frac{1}{h^d} k(\frac{x}{h})$

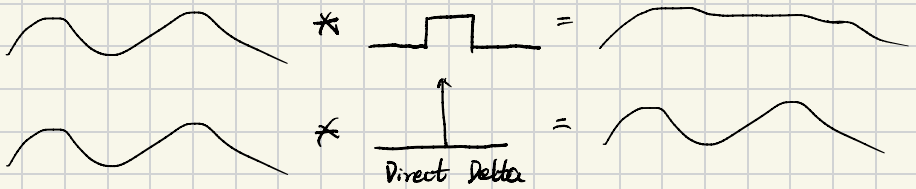
↑ scale amplitude; ↑ scale the width

$$\Rightarrow \hat{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \tilde{k}(x - x_i)$$

Mean: $\mathbb{E}[\hat{p}(x)] = \int p(u) \cdot \tilde{K}(x-u)\, du = p(x) * \tilde{k}(x)$

Convolution between the true p(x) and the kernel

blur the true p(x)



$$\delta(x) = \begin{cases} \infty, & x=0 \\ 0, & \text{other} \end{cases} \qquad \int f(x)\delta(x) = f(0)$$

only unbiased:

$$\tilde{k}(x) = \delta(x) = \lim_{h\to 0} \tilde{K}(x) \Rightarrow \hat{p}(x) = p(x)$$

Variance

$$\text{var}[\hat{p}(x)] \le \frac{1}{Nh^d} \left(\max_x k(x) \right) \mathbb{E}[\hat{p}(x)]$$

Not changable

For small variance, we need to have h large or N large

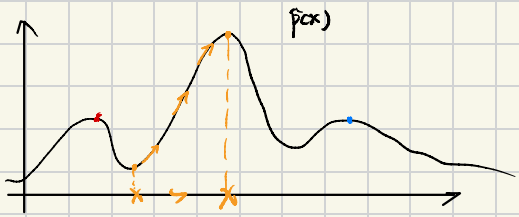
Summary: h controls the trade-off bias VS var.

$h \to 0 \Rightarrow$ bias $= 0$, var $= \infty$

$h \to \infty \Rightarrow$ bias $= \infty$ var $= 0$

Mean shift algorithm

Find a mode in $\hat{p}(x)$ (peak)



steps: 1) start a random point x_i

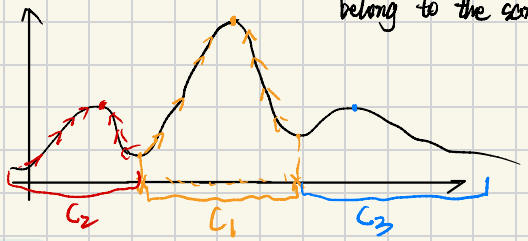
2) use grad. ascent to move uphill on $\hat{p}(x)$

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

3) repeat, $\hat{x}$ converge to a mode

repeat over all samples to find all modes

Clustering: the x_i that share the same mode. belong to the same cluster



radially symmetric kernel

$$K(x) = \alpha \bar{k}(\|x\|^2)$$

constant, kernel profile, distance

Gaussian $K(x) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|x\|^2}$, $\alpha = \frac{1}{(2\pi)^{d/2}}$

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

Density estimate: $\hat{p}(x) = \frac{\alpha}{Nh^d} \sum_{i=1}^{N} \bar{k}\left(\left\|\frac{x - x_i}{h}\right\|^2\right)$

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

$$\nabla \hat{p}(x) = \underbrace{\frac{2\alpha}{Nh^d}}_{c} \cdot \underbrace{\left(\sum_{i=1}^{N} \bar{g}\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \right)}_{\approx \text{density estimate using } \bar{g}(r)} \cdot \left(\frac{\sum_i x_i \bar{g}\left(\left\|\frac{x - x_i}{h}\right\|^2\right)}{\sum_i \bar{g}\left(\left\|\frac{x - x_i}{h}\right\|^2\right)} - x \right)$$

$\approx \hat{g}(x)$

numerator: weighted sum of x_i that are close to x, weighted by $\bar{g}(\cdot)$

denom: sum of weight

whole: weighted avg of points close to x

point

mean shift vector of x, $m(x)$

gradient ascent

$$\hat{x}^{(k+1)} = \hat{x}^{(k)} + \lambda \cdot \nabla_p(\hat{x}^{(k)})$$

$$= \hat{x}^{(k)} + \lambda \cdot c \cdot \hat{g}(\hat{x}^{(k)}) \cdot m(\hat{x}^{(k)})$$

λ: step size for gradient ascent

adaptive step size

$$\lambda = \frac{1}{\hat{g}(\hat{x}^{(k)}) \cdot c}$$

→ $\hat{g}(x)$ is small ⇒ distant ⇒ large step

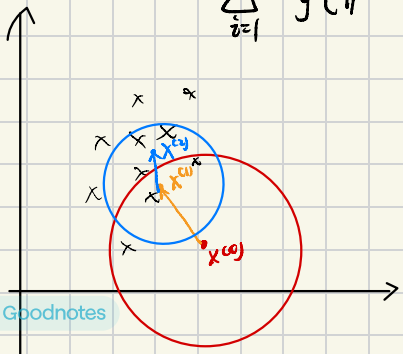
→ $\hat{g}(x)$ is large ⇒ close ⇒ small step

$$\hat{x}^{(k+1)} = \hat{x}^{(k)} + \frac{1}{\hat{g}(\hat{x}^{(k)}) \cdot c} \cdot c \cdot \hat{g}(\hat{x}^{(k)}) \cdot m(\hat{x}^{(k)})$$

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

$$= \frac{\sum_{i=1}^{N} x_i \, \bar{g}\left(\left\| \frac{x - x_i}{h} \right\|^2\right)}{\sum_{i=1}^{N} \bar{g}\left(\left\| \frac{x - x_i}{h} \right\|^2\right)}$$

mean-shift



Note: the algorithms is guaranteed to converge to a stationary point, i.e., find the mode, if the kernel profile is monotonously decrease and convex.