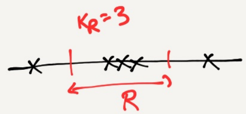
CS5487 Lecture Notes (2022B)
Prof. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

$\boxed{\text{Lecture 5}}$ Non-parametric Density Estimation

- so far we have seen parametric densities like Gaussian, GMM, etc, which make 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 $\{x_1, \ldots, x_n\}$

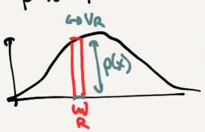


- Consider a region R
- Define $P = P(x \in R) = \int_{x \in R} p(x)\,dx$

  prob. of a point in R

- Define K_R = # points inside R
- Estimate of P: $\hat{P} = \frac{K_R}{n}$ ☆
- Assume R is small, then

☆ $\hat{P} \approx p(x) V_R$, V_R = volume of R, x = center of R



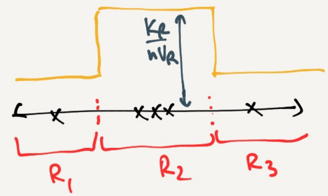
(approximate the integral over R with rectangle)

$$\underbrace{\hat{P} = \frac{K_R}{n}, \quad \hat{P} = p(x) V_R}$$

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

← # points in R

← volume of R



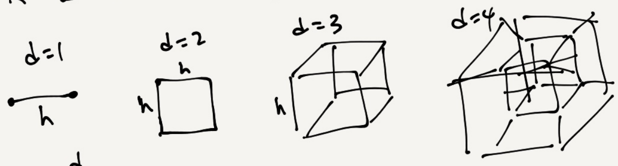
This is just a histogram, but we can extend it.

How to choose R?

1) keep V_R fixed, & let K_R vary. → Parzen windows ✓ Kernel density estimation
2) keep K_R fixed, & let V_R vary. → k-NN estimation. (see PRML)

Kernel Density Estimators

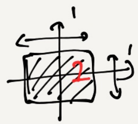
- let R be a d-dim. hypercube w/ side of h.



$V_R = h^d$

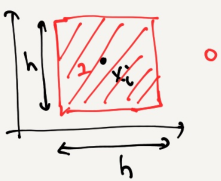
- introduce a window:

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

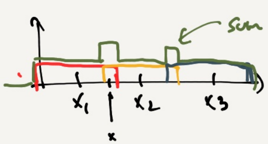


(Parzen window, kernel function)

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



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

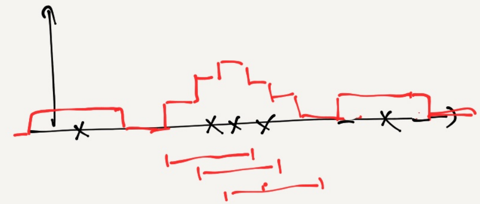


"stacking boxes centered at all x_i's."

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

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

KDE - estimation w/ interpolation btwn samples.



Other kernel functions

constraints: $k(x) \ge 0$, $\int k(x)\,dx = 1$ } i.e. a valid pdf.

Examples

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

unit sphere: $k(x) = \begin{cases} \frac{1}{c}, & \|x\|^2 \le 1 \\ 0, & \text{otherwise} \end{cases}$, c is volume of a unit sphere.

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

$$\hat{p}(x) = \frac{1}{nh^d} \sum_i k\left(\frac{x - x_i}{h}\right) = \overset{\pi_i}{\frac{1}{n}} \sum_i N(x \mid \underset{\text{mean } i\text{th comp.}}{x_i}, \underset{\text{covar.}}{h^2 I})$$

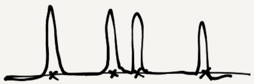
GMM w/ n components

Bandwidth parameter h

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

Intuitively,

h small



noisy estimate if not enough samples.

h large

blurry estimate if too many points.

Convergence Analysis

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

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

$\Rightarrow$ we can look at bias/variance.

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

② $\lim_{n\to\infty} var(\hat{p}(x)) = 0$

Define $\tilde{k}(x) = \underbrace{\frac{1}{h^d}}_{\text{scales the amplitude}} k\Big(\underbrace{\frac{x}{h}}_{\text{scales width of kernel}}\Big) \Rightarrow \hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \tilde{k}(x-x_i)$

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

$\vdots$

$= \int p(\mu)\tilde{k}(x-\mu)\,d\mu$

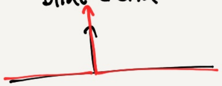
$= p(x) * \tilde{k}(x)$

convolution of true $p(x)$ w/ the kernel $\tilde{k}$ $\Rightarrow$ 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



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

Variance: $var(\hat{p}(x)) \le \frac{1}{nh^d}\max_x\big(k(x)\big)\, E[\hat{p}(x)]$

(see tutorial)

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

☆ h controls the tradeoff btwn bias & variance

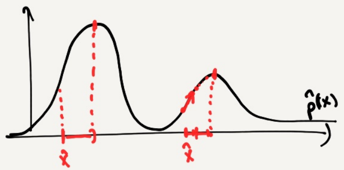
$\begin{cases} h\to 0 \Rightarrow \text{bias} = 0, & \text{var is large} \\ h\to\infty \Rightarrow \text{bias} \neq 0, & \text{var} = 0 \end{cases}$

How to select h?

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

Mean-Shift algorithm (Comaniciu & Meer)

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



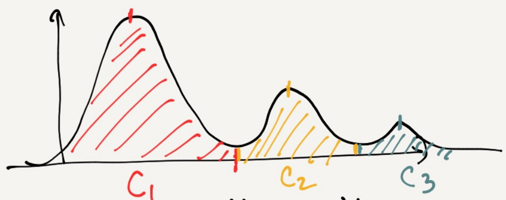
Idea:

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.



Consider radially symmetric kernels.

$$\underset{\text{kernel}}{k(x)} = \underset{\text{constant}}{\alpha}\, \underset{\text{kernel profile}}{\bar{k}}(\|x\|^2)$$

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

$$\bar{k}(r) = e^{-\frac{1}{2}r}, \quad \alpha = (2\pi)^{-d/2}$$

KDE

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

Define

$\bar{g}(r) = -\bar{k}'(r)$

Gaussian $\bar{g}(r) = \frac{1}{2} e^{-\frac{1}{2}r}$

gradient

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

"mean-shift" vector: diff. btwn mean inside a window & the center of window.

$= m(x)$

$$\nabla \hat{p}(x) = \frac{\alpha}{n \cdot h^d} \cdot -\sum_i \bar{g}\left(\left\|\frac{x - x_i}{h}\right\|^2\right) \cdot 2\left(\frac{x - x_i}{h}\right) \cdot \frac{1}{h}$$

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

Gradient Ascent

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

updated soln — current soln — step size is important for convergence.

Use an adaptive stepsize:

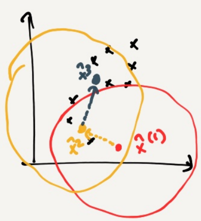
$$\lambda = \frac{1}{\hat{g}(x)}$$

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

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

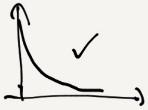
$$\Rightarrow \hat{x}^{(k+1)} = \hat{x}^{(k)} + \frac{1}{\cancel{g(\hat{x}^{(k)})}} \cancel{g(\hat{x}^{(k)})} m(\hat{x}^{(k)})$$

$$\boxed{\hat{x}^{(k+1)} = \frac{\sum_i x_i \bar{g}\left(\left\|\frac{\hat{x}^{(k)} - x_i}{h}\right\|^2\right)}{\bar{g}\left(\left\|\frac{\hat{x}^{(k)} - x_i}{h}\right\|^2\right)}}$$



mean-shift algorithm.

Note: • gavranteed to converge to a stationary point if kernel profile $k(r)$ is monotonically decrasing & convex.



Convolution

Def:

$$f(t) * g(t) = \int_0^t f(\tau) g(t-\tau)\, d\tau$$

properties:

$f * g = g * f$ (proof: change of variable $u = t - \tau$)

$f * (g * h) = (f * g) * h$

$f * (g + h) = f * g + f * h$