STAT 391

4/27/23

Lecture 10

- Gradient ascent
- Non-parametric density estimation (kernel d. e.)

LV non-parametric

Lecture Notes IV – Continuous distributions. Parametric density estimation.

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

CDF and PDF. Sampling

Examples of continuous distributions

ML estimation for continuous distributions

ML estimation by gradient ascent

Reading: Ch.5, 6

Examples of continuous distributions

$$\mathcal{F}_1 = \{u_{[a,b]},\ a < b\} \quad \text{uniform} \tag{5}$$

$$f(x; a, b) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$\mathcal{F}_2 = \{N(.; \mu, \sigma^2)\} \quad \text{normal} \tag{7}$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

$$F(x; a, b) = \frac{1}{1 + e^{-ax-b}}, \quad a > 0 \quad \text{logistic} \tag{9}$$

$$f(x; a, b) = \frac{ae^{-ax-b}}{(1 + e^{-ax-b})^2} \tag{10}$$

ML estimation by gradient ascent

Maximize $\ell(a,b)$ by gradient ascent

$$l(a,b) = n\ln a - a\sum_i x_i - nb - 2\sum_{i=1}^n \ln(1+e^{-ax_i-b})$$

$$\frac{\partial l}{\partial a} = \frac{n}{a} - \sum_{i=1}^n x_i \frac{1-e^{-ax_i-b}}{1+e^{-ax_i-b}} = 0$$

$$\frac{\partial l}{\partial b} = -\sum_{i=1}^n \frac{1-e^{-ax_i-b}}{1+e^{-ax_i-b}} = 0$$

Algorithm

Initialize a^0, b^0 $\quad a^0>0$

Until convergence $t=1,2,\ldots$

$$a^t \leftarrow a^{t-1} + \eta \frac{\partial \ell}{\partial a}(a^{t-1}, b^{t-1})$$

$$b^t \leftarrow b^{t-1} + \eta \frac{\partial \ell}{\partial b}(a^{t-1}, b^{t-1})$$

Output $a^t, b^t \equiv a^{ML}, b^{ML}$

Stopping G.A.

0) always have t_{max} limit

1) relative increase in ℓ safety

$$\frac{\ell(a^t,b^t) - \ell(a^{t-1},b^{t-1})}{\ell(a^{t-1},b^{t-1})} \leq \text{tolerance}$$

$10^{-10} < \text{tolerance} \sim \frac{1}{\sqrt{n}}$ $\quad \frac{1}{100}$

ML estimation by gradient ascent

Normalize $\ell(a,b)$!!

$$\ell^{norm}(a,b) \leftarrow \frac{\ell(a,b)}{n}$$

$$l(a,b) \;=\; n\ln a - a\sum_i x_i - nb - 2\sum_{i=1}^{n}\ln(1+e^{-ax_i-b})$$

$$\frac{\partial l}{\partial a} \;=\; \frac{n}{a} - \sum_{i=1}^{n} x_i \frac{1-e^{-ax_i-b}}{1+e^{-ax_i-b}} \;=\; 0$$

$$\frac{\partial l}{\partial b} \;=\; -\sum_{i=1}^{n} \frac{1-e^{-ax_i-b}}{1+e^{-ax_i-b}} \;=\; 0$$

3) $$\underbrace{\sqrt{\left(\frac{\partial \ell}{\partial a}\right)^2 + \left(\frac{\partial \ell}{\partial b}\right)^2}}_{\text{length of } \begin{bmatrix} \frac{\partial \ell}{\partial a} \\ \frac{\partial \ell}{\partial b} \end{bmatrix}} < \text{tol}$$

ML estimation by gradient ascent

$$l(a,b) = n \ln a - a \sum_i x_i - nb - 2 \sum_{i=1}^{n} \ln(1 + e^{-ax_i - b})$$

$$\frac{\partial l}{\partial a} = \frac{n}{a} - \sum_{i=1}^{n} x_i \frac{1 - e^{-ax_i - b}}{1 + e^{-ax_i - b}} = 0$$

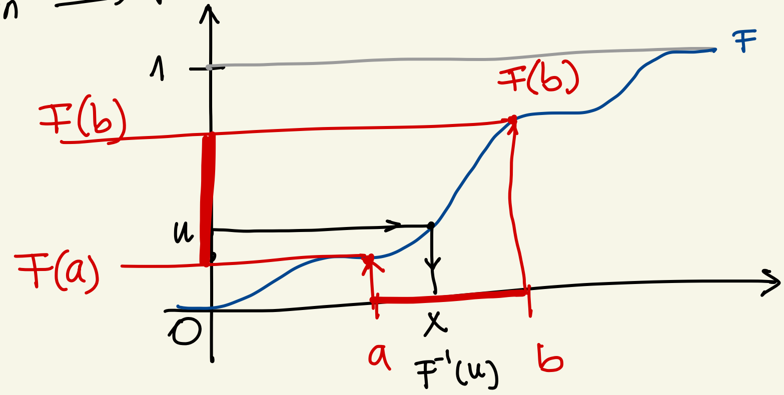
$$\frac{\partial l}{\partial b} = -\sum_{i=1}^{n} \frac{1 - e^{-ax_i - b}}{1 + e^{-ax_i - b}} = 0$$

Sampling from a continuous distribution

Given F CDF of P on $(-\infty, \infty)$

Wanted sample $x \sim P$ $\Leftrightarrow$ if many samples taken $\longrightarrow P$

1. Sample $u \sim U[0,1]$
   $u = \text{rand}();$
2. Output $x = F^{-1}(u)$



Why?

$$\Pr[a,b] = F(b) - F(a)$$

Alg

$$\Pr[x \in [a,b]]$$

$$x = F^{-1}(u) \Leftrightarrow F(x) = u$$

$$\Pr[x \in [a,b]] = \Pr[u \in [F(a), F(b)]] = F(b) - F(a)$$

definition of uniform [0,1]

Lecture Notes V – Non-parametric density estimation

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

April, 2023

S discrete ✓ ML

S continuous $(-\infty, \infty)$

models — parametric ✓ ML
— non-parametric ← NO ML !!

$\mathcal{F} = \{f_\theta\}$

Kernels

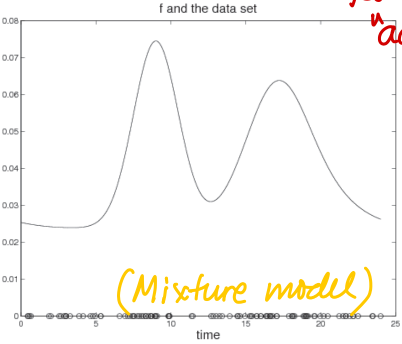
Kernel density estimators

Choosing h by Cross-Validation and the Bias-Variance trade-off

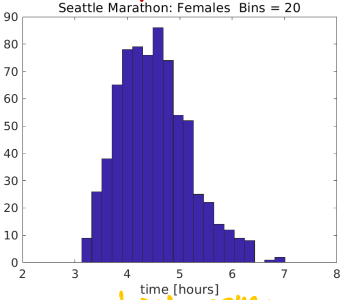
Reading: Ch. 7

Why non-parametric?

— fit any shape
"adapt" to shape of data



(Mixture model)



histogram

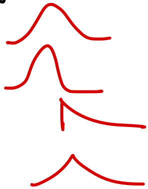
Parametric families

Normal

Logistic

Exponential

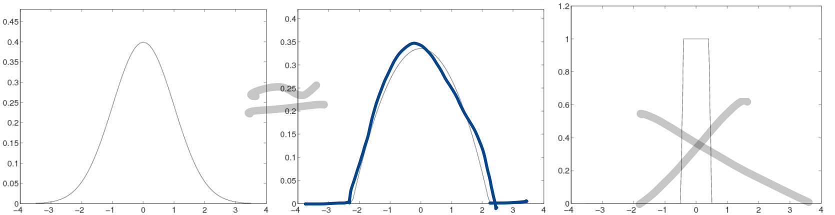
[Cauchy]



Kernel density estimation ←

Kernels

(Shape of kernel not important)



Gaussian kernel $N(0,1)$ — Epanechnikov — Square unif $[-0.5, 0.5]$

Kernel = simple density $\to$ $k(z) \geq 0$, $\int_{-\infty}^{\infty} k(z) = 1$

$k(z)$

$k : (-\infty, \infty) \to [0, \infty)$

1. $k(-z) = k(z)$

2. $k(z) \searrow$ for $z > 0$

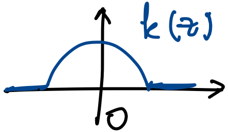
[3.] $\int_{-\infty}^{\infty} z^2 k(z)\, dz = 1$ standardization

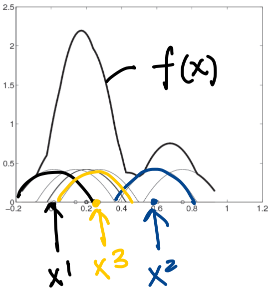
Kernel density estimators

$$f(\cancel{x}) = \underbrace{\frac{1}{n}}_{\text{normalization}} \sum_{i=1}^{n} k(\cancel{x} - \underbrace{x^i}_{\text{data}})$$

$$k(\underbrace{x - x^i}_{z})$$

shift to x^i





$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{n}\sum_i \underbrace{\int_{-\infty}^{\infty} k(x-x^i)\,dx}_{1}$$

$$= \frac{1}{n}\cdot n = 1$$

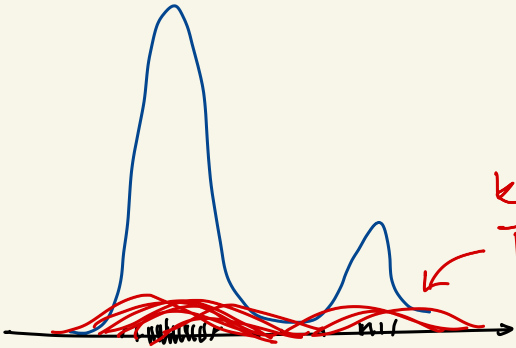
$\Downarrow$

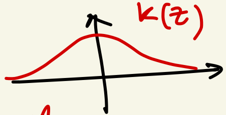
f is a density

(bandwidth)
$\boxed{h = \text{kernel width}}$

$$k_h(z) = \frac{1}{h} k\left(\frac{z}{h}\right)$$

unit on x axis (z)





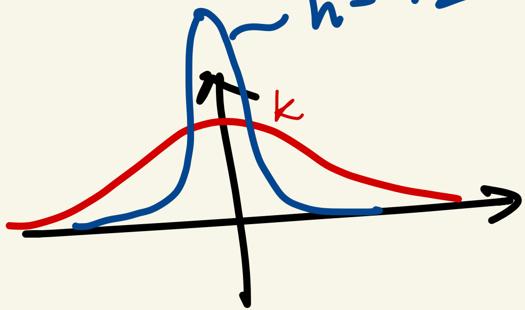
kernel

Too flat!! $\Rightarrow$ scale z

Exercise:
$$\int_{-\infty}^{\infty} k_h(z)\, dz = 1$$

$$f'(x) = \frac{1}{n} \sum_i k'(x - x^i) \Rightarrow |f'| \leq \frac{1}{n} \sum_{i=1}^{n} a = a$$

$$|k'(z)| < a$$



$$h = \frac{1}{2}$$

Kernel density estimators

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - x^i}{h}\right)$$

data

width

smoothness par