Manifold Learning

Density Estimation

Jairo Cugliari

Master Informatique

Parcours Data Mining

Université Lumière Lyon 2

Université de Lyon

Univariate data representation

Old Faithful Geyser Data : waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

- Data : 272 obs $\times$ 2 vars
- Methods to analyze this data : summaries, plots, smth more clever ?

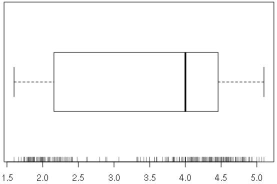
	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85

Graphical representation of univariate data

Stem-and-fleaf

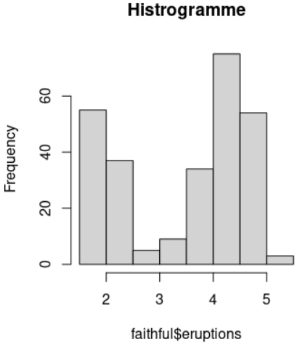
boxplox

Histogramme




```
The decimal point is 1 digit(s) to the left of the |

16 | 070355555588
18 | 000022233333335577777777888822335777888
20 | 00002223378800035778
22 | 0002335578023578
24 | 00228
26 | 23
28 | 080
30 | 7
32 | 2337
34 | 250077
36 | 0000823577
38 | 2333335582225577
40 | 00000033577888880022335555577778
42 | 033355555778800233333555577778
44 | 02222335557780000000023333357778888
46 | 000023335770000023578
48 | 00000022335800333
50 | 0370
```



Univariate Density Estimation

- Data : $X_1, \ldots, X_n$ iid real random variables
- X has (unknown) density $f(x)$
- Goal: estimate f making mild assumptions
  - parametric : assume f belong to a parametric family and guess θ), e.g. $\mathcal{N}(\mu, \sigma)$
  - nonparametric : assume just some kind of regularity (i.e. smoothness)
- Non parametric estimators are very popular
  - histograms (H)
  - frequency polygons (FP)
  - kernel density estimators (KDE)

Learning the whole distribution is feasible

Recall : the empirical cumulative distribution function on n samples is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

- The Glivenko-Cantelli theorem says

$$\max_x |\hat{F}_n(x) - F(x)| \longrightarrow 0$$

Intuitively, the empirical CDF converges to the true CDF everywhere, i.e. the maximum gap between the two of them goes to zero

Can we use the empirical CDF to estimate a density?

- Yes, but it's discrete and doesn't estimate a density well
- How to put non-zero density between observations ?
- If a random variable X has probability density f, then

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

- Thus, a naive estimator would be

$$\widehat{f}(x) = \frac{1}{2nh} [\# \text{ of } x_i \text{ falling in } (x - h, x + h)]$$

Naive estimator

- Or, equivalently

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}w\left(\frac{x-x_i}{h}\right)$$

where w is a weight function defined as

$$w(x) = \begin{cases} 1/2 & |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

- In short, a naive estimate is constructed by placing a box of width $2h$ and height $\frac{1}{2nh}$ on each observation, then summing to obtain the estimate

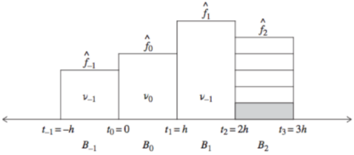
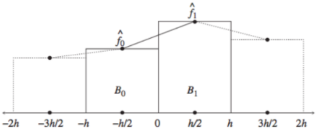
Histrograms

- Simple but powerful to estimate distributions f.
- Formally, split the sample space up to bins $(B_1, B_2, \ldots, B_m)$, and count the absoluty frequency on each bin $(\nu_1, \nu_2, \ldots, \nu_m)$.
- The Histrogram estimator $f_n^H(x)$ is defined as

$$\hat{f}_n^H(x) = \frac{\nu_j}{nh}, \qquad x \in B_j, h = 1/m.$$

- If we hold the bins fixed and take more and more data, then the relative frequency for each bin will converge on the bin's probability : $\hat{f}_n^H(x) \mapsto f(x)$

Non parametric Density Estimation

$\hat{f}_n(x) = \frac{\nu_j}{nh} = \frac{1}{nh}\sum_{i=1}^{n} \mathbb{1}_{[t_j, t_{j+1}]}(x_i)$	
$\hat{f}_n(x) = \left(\frac{1}{2} - \frac{x}{h}\right)\hat{f}_0 + \left(\frac{1}{2} + \frac{x}{h}\right)\hat{f}_1$	
$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$	

We'll study the KDE.

We need :

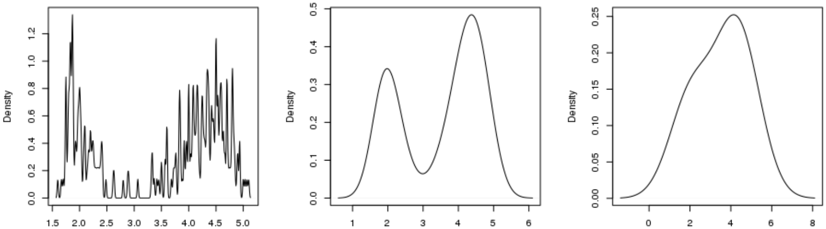
- a kernel K function bounded 2nd moment)
- a positive number h called the bandwidth

Kernel Density Estimation (KDE)

- The KDE of f is defined as

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

- Several kernel functions exists
- The crucial quantity is h which must be correctly tuned



MSE analysis of KDE

- One popular loss function is the mean-square-error

$$L(f(x), \widehat{f_n}(x)) = \mathbb{E}[\widehat{f_n}(x) - f(x)]^2$$

- The MSE can be expressed as

$$L(f, \widehat{f_n}) = \text{bias}^2[\widehat{f_n}] + \text{Var}[\widehat{f_n}]$$

KDE calibration

How to choose the optimal value h^* ?

- Normal reference : if f and K are normal, $h^* = 1.06\sigma n^{-1/5}$
  - Estimate σ by $\hat{\sigma} = \{s, \mathrm{IQR}/1.34\}$, where s is the empirical standard deviation and IQR the interquartile range
  - Use $h^* = 1.06\hat{\sigma} n^{-1/5}$
- Cross validation
  - CV score function $\hat{J}(h) = \int \hat{f}^2(x)dx - (2/n)\sum_{i=1}^{n} \hat{f}_{-i}(X_i)$
  - Use $h^* = \arg\min \hat{J}(h)$