Manifold Learning

Density Estimation

Jairo Cugliari

Master Informatique

Parcours Data Mining

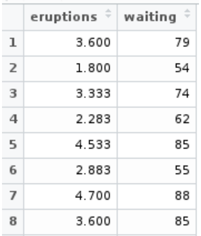
Université Lumière Lyon 2

Université de Lyon

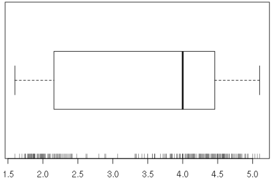
Density estimation

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 × 2 vars
- ○ Methods to analyze this data : summaries, plots, smth cleverer?



	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



Density Estimation

- Data : $X_1, \ldots, X_n$ from an unknown density f
- Goal: estimate f making mild assumptions
  - nonparametric vs parametric
- Histograms are a popular choice but ...

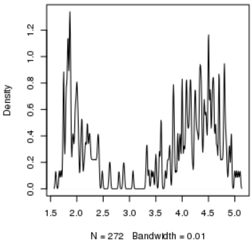
- We'll study the kernel density estimator. We need :
  - a kernel K function (centred prob mass function with bounded 2nd moment)
  - a positive number h called the bandwidth

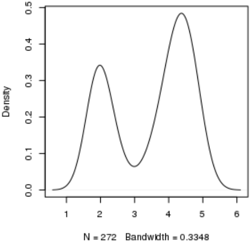
Kernel Density Estimation (KDE)

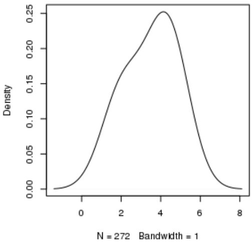
- The KDE of f is defined as

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

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







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 $\hat{\sigma} = \{s, \text{IQR}/1.34\}$, 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 $h^* = \arg\min \hat{J}(h)$