Unsupervised Data Mining: From Batch to Stream Mining Algorithms

Prof. Dr. Stefan Kramer

Johannes Gutenberg University Mainz

Outline

- Interval-based methods
- Kernel density estimation
- Tree-based methods

Acknowledgements

- J. Siekiera

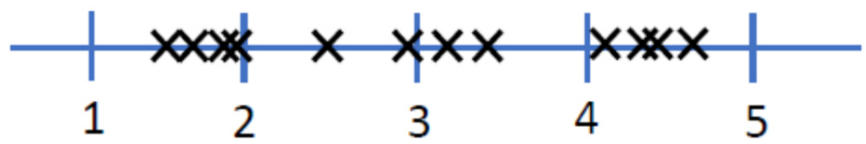
Histogram Estimators

Motivation

- Given: lengths of eruptions of Old Faithful Geysir in min
- Goal: How are data distributed?
- Assumptions: Data are governed by some probability distribution
- Use density to describe the distribution of the data

```
[4.37, 3.87, 4.00, 4.03, 3.50, 4.08, 2.25,
 4.70, 1.73, 4.93, 1.73, 4.62, 3.43, 4.25,
 1.68, 3.92, 3.68, 3.10, 4.03, 1.77, 4.08,
 1.75, 3.20, 1.85, 4.62, 1.97, 4.50, 3.92,
 4.35, 2.33, 3.83, 1.88, 4.60, 1.80, 4.73,
 1.77, 4.57, 1.85, 3.52, 4.00, 3.70, 3.72,
 4.25, 3.58, 3.80, 3.77, 3.75, 2.50, 4.50,]
```

Eruptionslängen des Old Faithful Geysir



Density

- X: real-valued random variable and a, b ∈ R
- f(x): density of X
- It must hold that $\int_{-\infty}^{\infty} f(x)dx = 1$
- $P(a < X < b) = \int_{a}^{b} f(x)dx = F(b) - F(a)$

Histograms

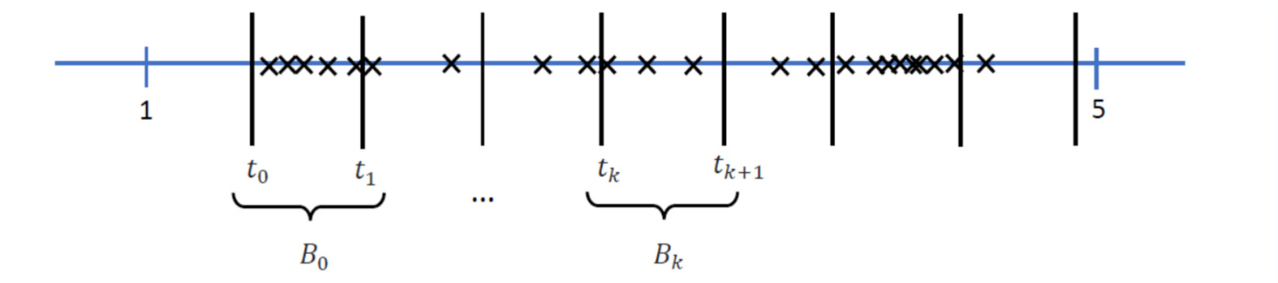
N: number of instances in total

$B_k = [t_k, t_{k+1})$: interval with k in N_0

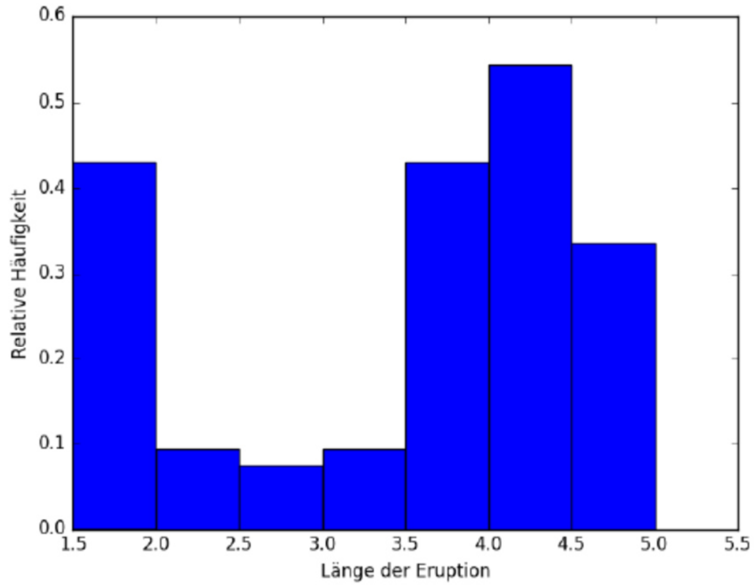
N_k: number of instances in interval B_k

$$\hat{f}(x) = \frac{n_k}{N(t_{k+1} - t_k)}$$ for x in B_k

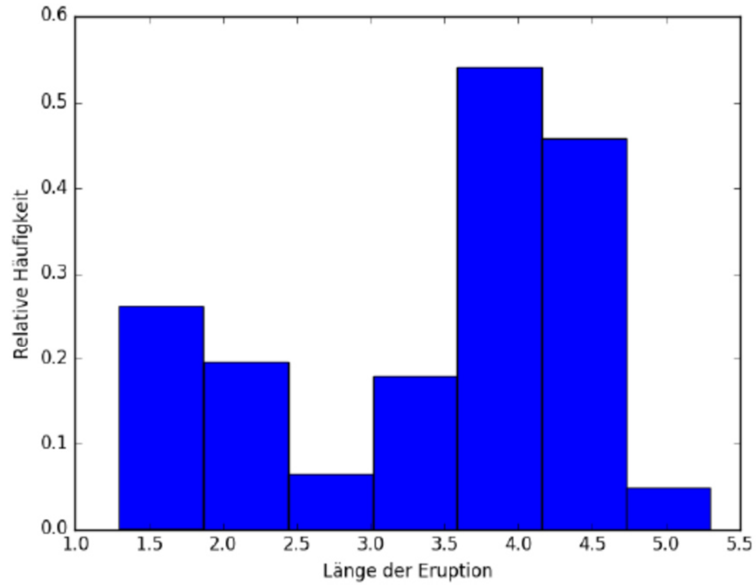
If interval width fixed, set $(t_{k+1} - t_k) = h$



Choice of Origin t_0 Affects Result



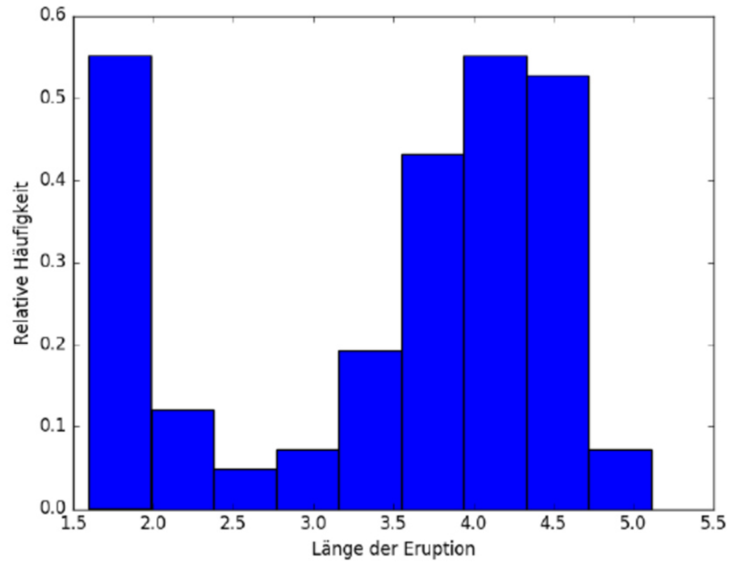
$\mathrm{h} = 0.5 \quad t_0 = 1.5$



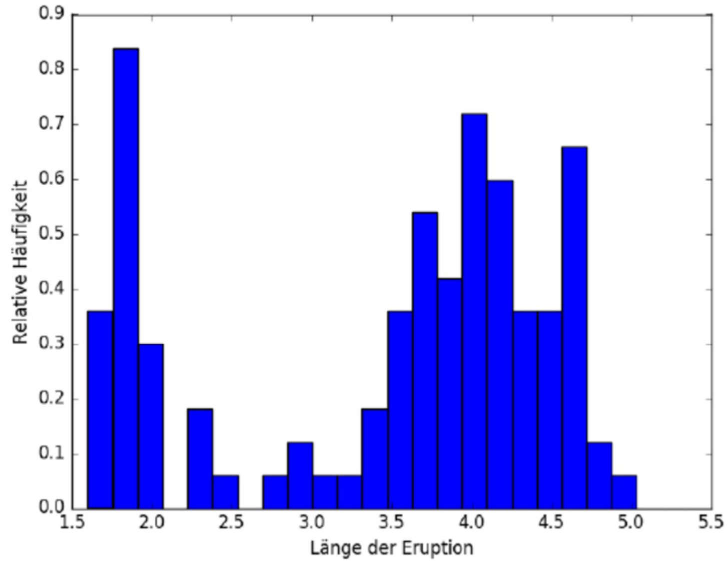
$\mathrm{h} = 0.5 \quad t_0 = 1.3$

Choice of Interval Width h Affects Results

h determines level of granularity



$\mathrm{h} = 0.4 \qquad t_0 = 1.6$



$\mathrm{h} = 0.17 \qquad t_0 = 1.6$

One has to avoid too general structure and overfitting.
Considerations about optimal h.

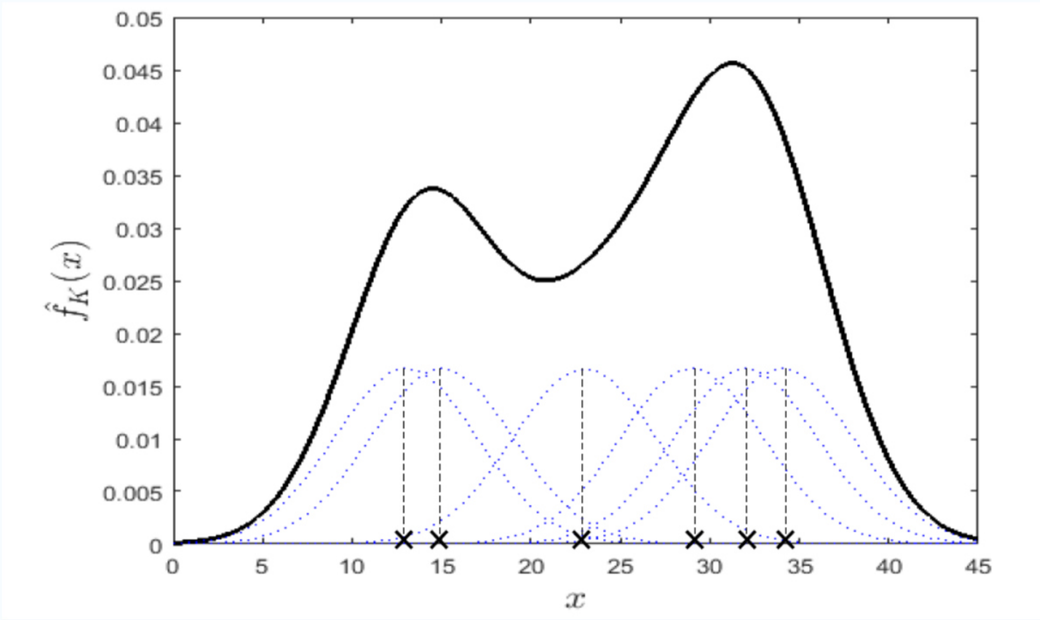
Summary Histograms

- Good for the presentation of one-dimensional data
- Fast calculation
- No continuous function obtained
- Dependency on the position of the origin

Kernel Density Estimators

Kernel Density Estimator (KDE)

- Place each instance into the center of a function
- Choose a continuous density K
- Sum up all individual functions

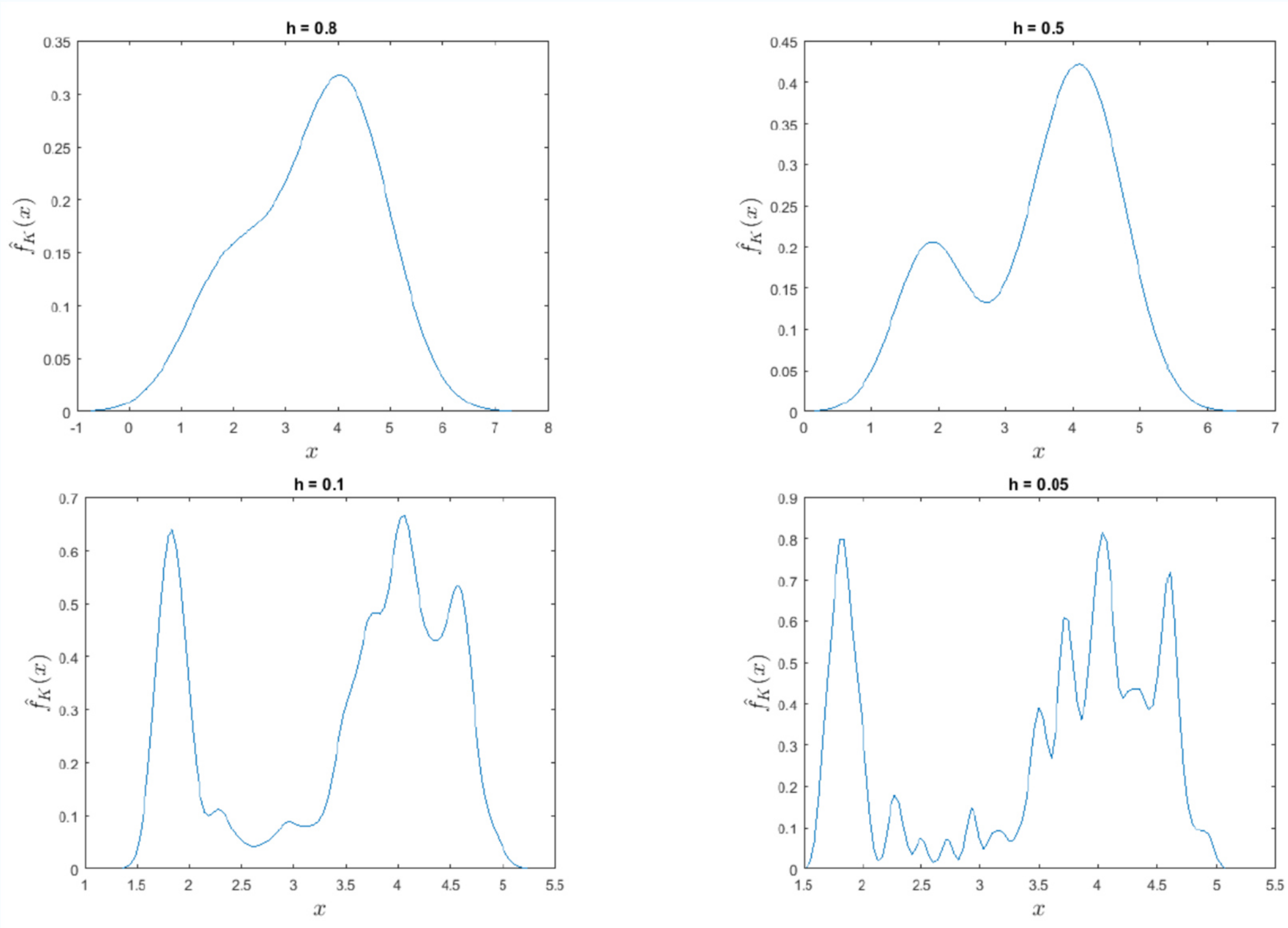


Kernel Density Estimators

$$\hat{f}_K(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \quad \text{with} \quad \int_{-\infty}^{\infty} K(x)dx = 1$$

- K(x) defines the form of the densities
- h defines the width
- $\hat{f}_K(x)$ inherits continuity and differentiability

Observations



Problem

- In regions with fewer data
  - higher noise
  - larger h required
- In regions with more data
  - less noise
  - smaller h required

Variable Kernel Method

- $d_{i,k}=|x_i - x_k|$: distance of x_i to the k-next point x_k

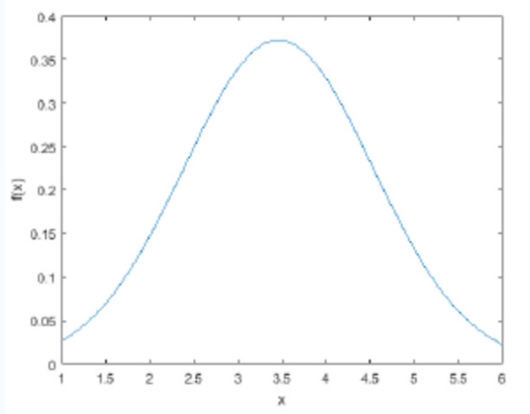
$$\hat{f}_K(x) = \frac{1}{Nh}\sum_{i=1}^{N}\frac{1}{d_{i,k}}K\left(\frac{x-x_i}{hd_{i,k}}\right)$$

- Width of kernel depends on distance of given data
- K determines contribution of local densities

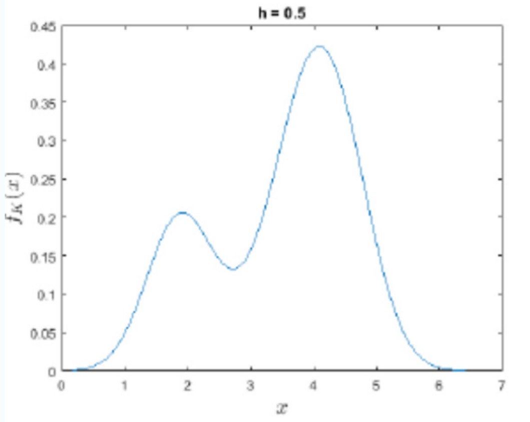
- Advantage: very accurate estimator
- Disadvantage: computationally intensive

Categorization of Models

- **Parametric models**
  - probability distributions assumed
  - model specified except for parameters
  - strongly dependent on model assumptions



- **Non-parametric models**
  - Example: kernel density estimation
  - Independent of assumed distribution



- **Semi-parametric models**

Density Trees

Density Trees

- Idea: decision tree as a basis for an estimator
- Analogously to classification and regression trees
- **Advantages:**
  - automatic feature selection
  - processing of heterogeneous data
  - Interpretability

Let

- l: leaf of a tree
- V_l: minimal d-dimensional volume of leaf l

Then:

$$\hat{f}(x) = \frac{|l|}{NV_l}$$