Density Estimation

1

- Parametric techniques
  - Maximum Likelihood
  - Maximum A Posteriori
  - Bayesian Inference
  - **G**aussian **M**ixture **M**odels (GMM)
    - EM-Algorithm

- **Non-parametric techniques**
  - Histogram
  - Parzen Windows
  - k-nearest-neighbor rule

Non-parametric Techniques

2

- Common parametric forms rarely fit the densities encountered in practice.

- Classical parametric densities are unimodal, whereas many practical problems involve multimodal densities.

- Non-parametric procedures can be used with arbitrary distributions and without the assumption that the form of the underlying densities are known.

Histograms

- Conceptually most simple and intuitive method to estimate a p.d.f. is a histogram.
- The range of each dimension x_i of vector $\boldsymbol{x}$ is divided into a fixed number m of intervals.
- The resulting M boxes (bins) of identical volume V count the number of points falling into each bin:
- Assume we have N samples ($\boldsymbol{x_i}$) and the number of points $\boldsymbol{x}_l$ in the *j-th* bin, b_j, is k_j. Then the histogram estimate of the density is:

$$p(x) = \frac{k_j / N}{V}, \quad x \in b_j$$

Histograms

$p(x)$

- … is constant over every bin b_j
- … is a density function

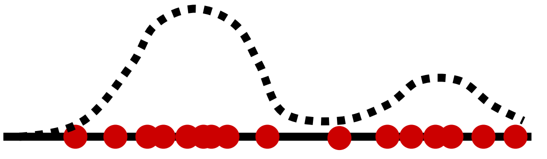
$$\int p(\boldsymbol{x})dx = \sum_{j=1}^{M} \int_{b_j} \frac{k_j}{NV} dx = \frac{1}{N} \sum_{j=1}^{M} k_j = 1$$

- The number of bins M and their starting positions are "parameters". However only the choice of M is critical. It plays the role of a smoothing parameter.

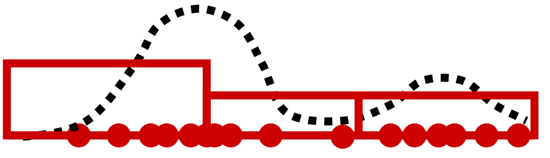
Histograms: Example

5

- Assume one dimensional data sampled from a combination of two Gaussians



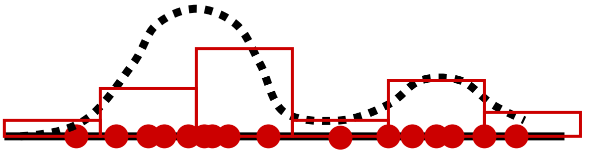
- 3 bins



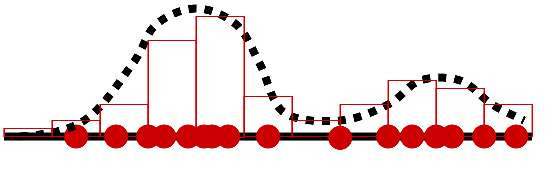
Histograms: Example

6

- 7 bins



- 11 bins



Histogram Approach

- Histogram p.d.f. estimator is very efficient since it can be computed online (only update counters, no need to keep all data)

- Usefulness is limited to low dimensional vectors, since number of bins, M, grows exponentially with data's dimensionality d:

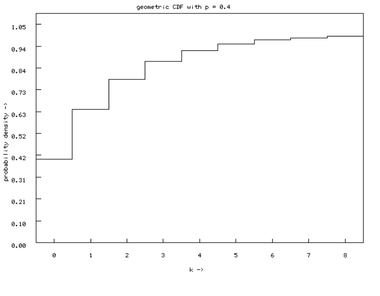
$$M = m^d$$

  - <u>"Curse of dimensionality"</u>

Parzen Windows: Motivation

- Consider set of 1-D samples $\{x_1, \ldots, x_N\}$ of which we want to estimate the density
- We can easily get estimate of cumulative distribution function (CDF) as:

$$P(x) = \frac{\#(\text{samples}) \leq x}{N}$$



- Density $p(x)$ is the derivative of the CDF
- But that is discontinuous !!

Parzen Windows:

- What we can do, is to estimate the density as:

$$p(x) = \frac{P(x + h/2) - P(x - h/2)}{h}, \qquad h > 0$$

- This is the proportion of observations falling within the interval $[x\text{-}h/2, x\text{+}h/2]$ divided by h .
- We can rewrite the estimate (already for d dim.):

$$p(\boldsymbol{x}) = \frac{1}{Nh^d} \sum_{i=1}^{N} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{h}\right)$$

with $$K(\boldsymbol{z}) = \begin{cases} 1 & |\boldsymbol{z}_j| \leq 1/2 \quad \forall\ j = 1 \ldots d \\ 0 & otherwise \end{cases}$$

Parzen Windows:

- The resulting density estimate itself is not continuous.
- This is because points within a distance $h/2$ of x contribute a value $1/N$ to the density and points further away a value of zero.
- Idea to overcome this limitation:

  Generalize the estimator by using a smoother weighting function (e.g. one that decreases as $|z|$ increases).

  This weighting function K is termed ***kernel*** and the parameter h is the ***spread*** (or ***bandwidth***).

Parzen Windows

11

- The kernel is used for interpolation: each sample contributes to the estimate according to its distance from x
- For a density, $p(\boldsymbol{x})$ must:
  - Be non-negative
  - Integrate to 1
- This can be assured by requiring the kernel itself to fulfill the requirements of a density function, ie.:

$$K(x) \geq 0 \quad \text{and} \quad \int K(\boldsymbol{z})\, d\boldsymbol{z} = 1$$

Parzen Windows: Kernels

12

Discontinuous Kernel Functions:

Rectangular:

$$K(\boldsymbol{x}) = \begin{cases} 0 & |\boldsymbol{x}| > 1/2 \\ 1 & |\boldsymbol{x}| \leq 1/2 \end{cases}$$

Triangular:

$$K(\boldsymbol{x}) = \begin{cases} 0 & |\boldsymbol{x}| > 1 \\ 1 - |\boldsymbol{x}| & |\boldsymbol{x}| \leq 1 \end{cases}$$

Smooth Kernels:

Normal:

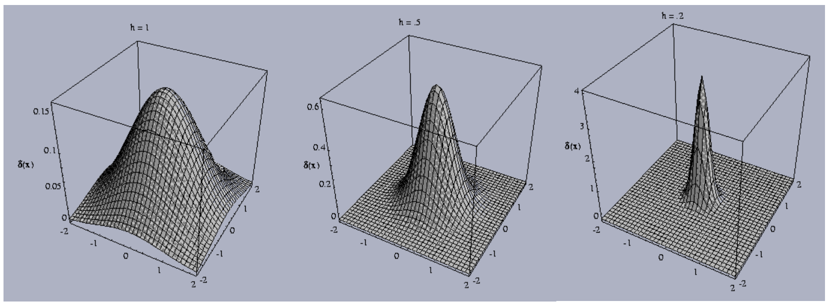
$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right)$$

Multivariate normal:
(radially symm. univ. Gaussian)

$$K(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(\frac{-\boldsymbol{x}^T \boldsymbol{x}}{2}\right)$$

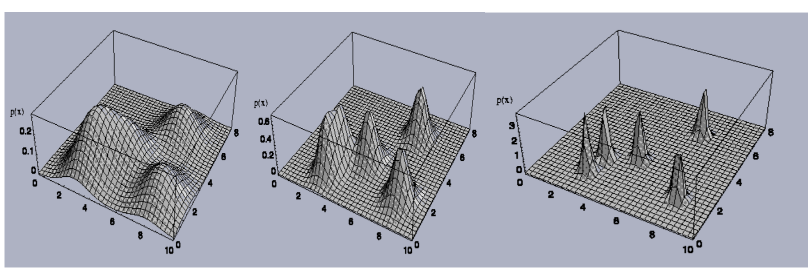
Parzen Windows: Bandwidth

- The choice of bandwidth is critical !



Examples of two-dimensional circularly symmetric normal Parzen windows for 3 different values of h.

Parzen Windows: Bandwidth

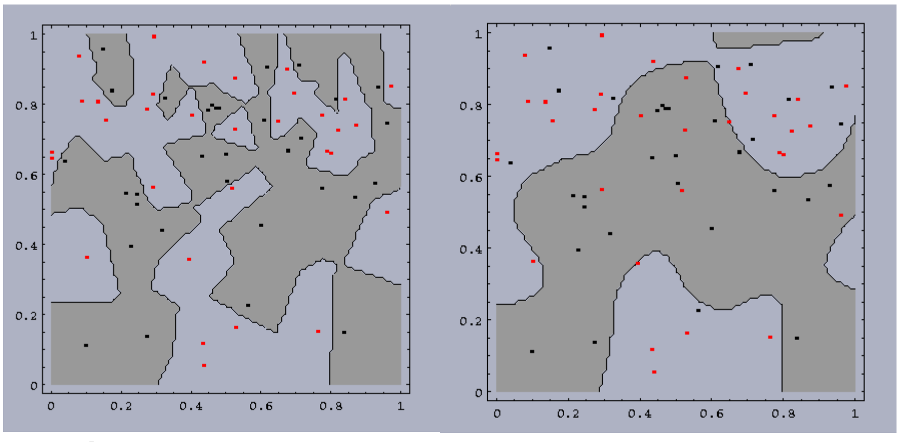


3 Parzen-window density estimates based on the same set of 5 samples, using windows from previous figure

- If h is too large the estimate will suffer from too little resolution.
- If h is too small the estimate will suffer from too much statistical variability.

Parzen Windows: Bandwidth

- The decision regions of a PW-classifier also depend on bandwidth (and of course of kernel).



Small h : more complicated boundaries

Large h : less complicated boundaries

k-Nearest-Neighbor Estimation

- Similar to histogram approach.

- Estimate $p(x)$ from N training samples by centering a volume V around x and letting it grow until it captures k samples.

- These samples are the k nearest neighbors of x.

- In regions of high density (around x) the volume will be relatively small.

- k plays a similar role as the bandwidth parameter in PW.

17

- Let N be the total number of samples and V the volume around x which contains k samples then

$$p(\boldsymbol{x}) = \frac{k}{N \cdot V(x)}$$

k-NN Decision Rule (Classifier)

- Suppose that in the k samples we find k_m from class ω_m (so that $\sum_{m=1}^{M} k_m = k$).

- Let the total number of samples in class ω_m be n_m (so that $\sum_{m=1}^{M} n_m = N$) .

18

k-NN Decision Rule (Classifier)

- Then we may estimate the class-conditional density $p(\boldsymbol{x}|\omega_m)$ as

$$p(\boldsymbol{x} \mid \omega_{\boldsymbol{m}}) = \frac{k_m}{n_m V}$$

and the prior probability $p(\omega_{\boldsymbol{m}})$ as

$$p(\omega_{\boldsymbol{m}}) = \frac{n_m}{N}$$

- Using these estimates the decision rule:

assign $\boldsymbol{x}$ to ω_m if $\quad \forall i: \quad p(\omega_{\boldsymbol{m}} \mid \boldsymbol{x}) \geq p(\omega_{\boldsymbol{i}} \mid \boldsymbol{x})$

translates (Bayes' theorem) to:

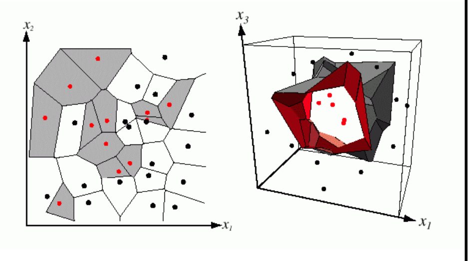
assign $\boldsymbol{x}$ to ω_m if $\quad \forall i: \quad k_m \geq k_i$

k-NN Decision Rule (Classifier)

19

- The decision rule is to assign $\boldsymbol{x}$ to the class that receives the largest vote amongst the k nearest neighbors of all classes M.

- For k=1 this is the nearest neighbor rule producing a Voronoi-tesselation of the training space.



- This rule is sub-optimal, but when the number of prototypes is large, its error is never worse than twice the Bayes error classification probability P_B.

$$P_B \leq P_{kNN} \leq P_B(2 - \frac{M}{M-1} P_B) \leq 2P_B$$

Non-parametric comparison

20

- Parzen window estimates require storage of all observations and n evaluations of the kernel function for each estimate, which is computationally expensive!

- Nearest neighbor requires the storage of all the observations.

- Histogram estimates do not require storage for all the observations, they require storage for the description of the bins. But for simple histograms the number of the bins grows exponentially with the dimension of the observation space.

Non-parametric Techniques

21

Advantages

- Generality: same procedure for unimodal, normal and bimodal mixture.
- No assumption about the distribution required ahead of time.
- With enough samples we can converge to an arbitrarily complicated target density.

Non-parametric Techniques

22

Disadvantages

- Number of required samples may be very large (much larger than would be required if we knew the form of the unknown density) .
- Curse of dimensionality.
- In case of PW and KNN computationally expensive (storage & processing).
- Sensitivity to choice of bin size, bandwidth,…