Non-Parametric Density Estimation

Sadegh Eskandari
Department of Computer Science, University of Guilan

Remember

- **Density Estimation**: given a finite set $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of observations for a random variable $\mathbf{x}$, the goal is to model the probability distribution $p(\mathbf{x})$.
- We will assume that the data points are independent and identically distributed (iid).

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x_n})$$

Density Estimation

- **Parametric**
  - Selecting a common distribution and estimating the parameters for the density function from the data
  - binomial and multinomial distributions for discrete random variables
  - Gaussian distribution for continuous random variables.
  - Parameter estimation procedure: maximum likelihood, Bayesian method
- **Non-Parametric**
  - Histograms, Nearest-Neighbours, Kernels

Non-Parametric Density Estimation

- We discussed probability distributions having specific functional forms governed by a small number of parameters whose values are to be determined from a data set.
- This is called the parametric approach to density modelling or density estimation.
- **Limitation**: the chosen density might be a poor model of the distribution that generates the data, which can result in poor predictive performance.
  - For example, if the process that generates the data is multimodal, then this aspect of the distribution can never be captured by a Gaussian, which is necessarily unimodal.
- Here, we consider nonparametric approaches to density estimation that make few assumptions about the form of the distribution.

Histograms

- We focus on the case of a single continuous variable x.
- Standard histograms simply partition x into distinct bins of width Δ and then count the number n_i of observations of x falling in bin i.
- The probability value for each bin is given by:

$$p_i = \frac{n_i}{N\Delta}$$

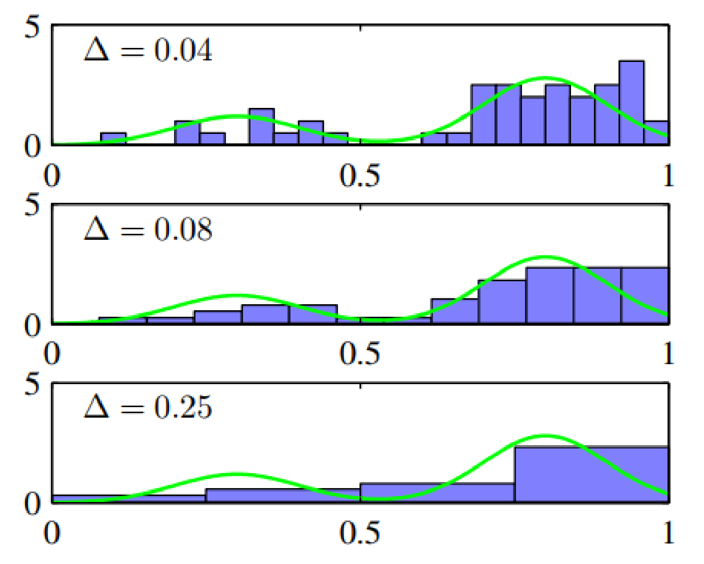
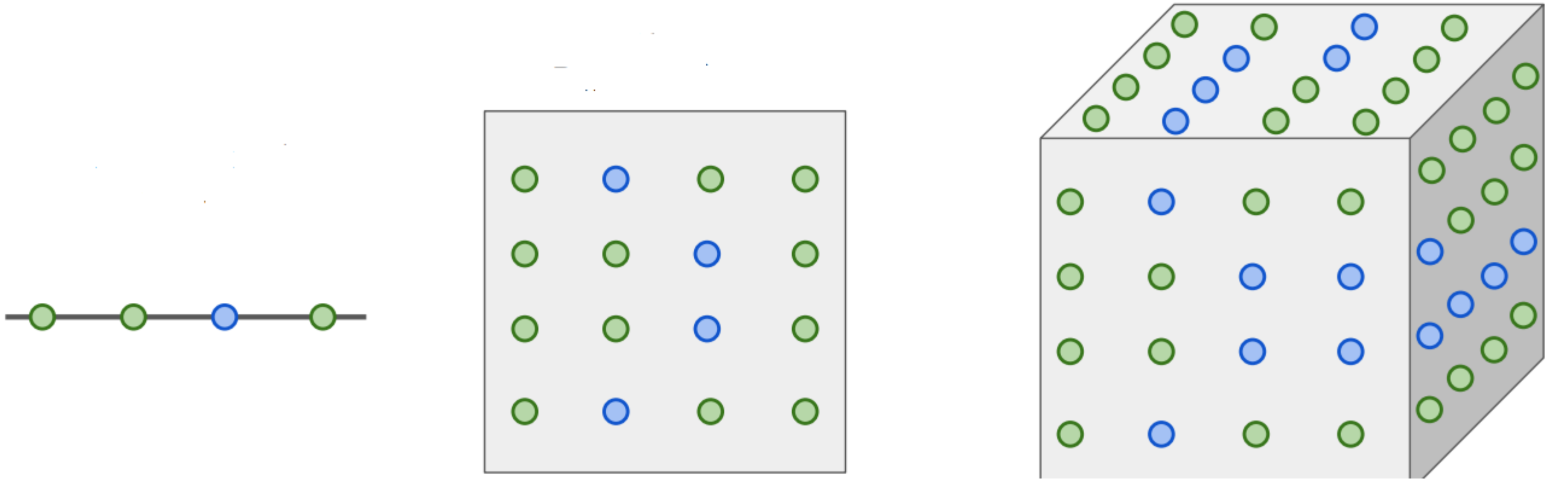


Figure: Three examples of density estimation corresponding to three different choices of the bin width

- ❑ Data (50 observations) is drawn from a mixture of two Gaussians (Green curve)
- ❑ Small Δ, spiky density model with structure not in the distribution
- ❑ Large Δ, smooth density model without underlying bi-modality
- ❑ Best results from intermediate Δ

Histograms

- **Limitations of the histogram approach**
  - Discontinues that are due to the bin edges
  - If we divide each variable in a D-dimensional space into M bins, then the total number of bins will be M^D (Curse of dimensionality)

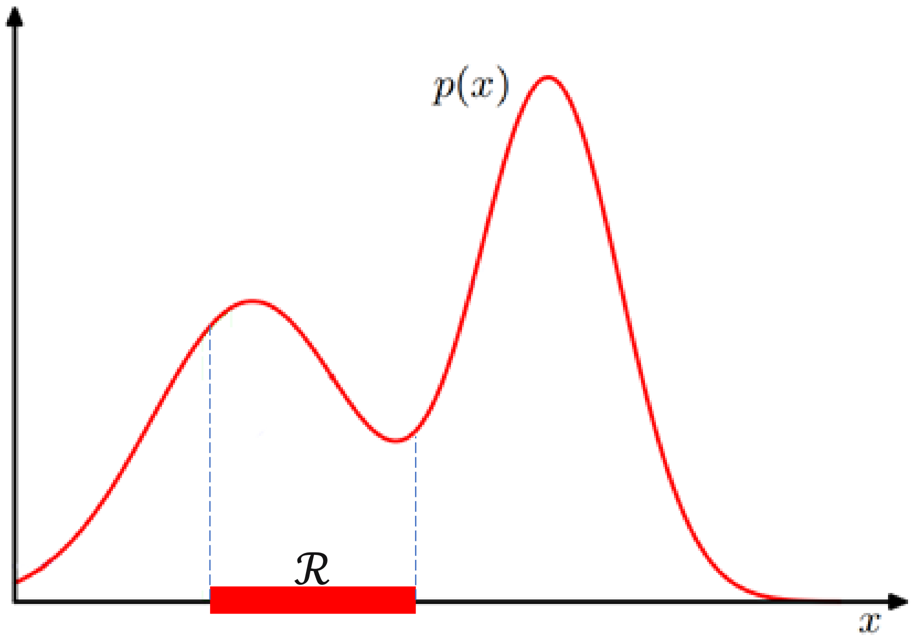


[Image from Stanford csn231n slides]

Kernel Density Estimators

- Let us suppose that observations are being drawn from some unknown probability density $p(\mathbf{x})$ in some D-dimensional space, and we wish to estimate the value of $p(\mathbf{x})$.
- Let us consider a small region $\mathcal{R}$ contaning $\mathbf{x}$.
- The probability mass associated with this region is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$$



- Now suppose that we have collected a data set comprising N observations drawn from $p(\mathbf{x})$.
  - Each data point has a probability P of falling within $\mathcal{R}$
  - Then for large N, the total number of points that lie inside $\mathcal{R}$ will be

$$K \simeq NP$$

(*)

Kernel Density Estimators

- If we also assume that the region $\mathcal{R}$ is sufficiently small that $p(\mathbf{x})$ is roughly constant over the region, then we have

$$P \simeq p(\mathbf{x})V$$

(**)

Where V is the volume of $\mathcal{R}$

- Combining (*) and (**), we obtain our density estimate in the form

$$p(\mathbf{x}) = \frac{K}{NV}$$

- Either
  - We can fix V and determine K from the data (kernel density estimation approach)
  - Or can fix K and determine V from the data (K-nearest neighbor approach)

Kernel Density Estimators

- To start with we take the region $\mathcal{R}$ to be a small hypercube centered on the point $\mathbf{x}$ at which we wish to determine the probability density.

- To count the number K of points falling within $\mathcal{R}$, define the following function

$$k(\mathbf{u}) = \begin{cases} 1, & |u_i| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise} \end{cases}$$

- The function $k(\mathbf{u})$ is an example of a kernel function, and in this context is also called a Parzen window.

- The quantity $k(\frac{\mathbf{x}-\mathbf{x}_n}{h})$ will be one if the data point $\mathbf{x}_n$ lies inside a cube of side h centered on $\mathbf{x}$, and zero otherwise.
- The total number of data points lying inside this cube will be

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

Kernel Density Estimators

$$p(\mathbf{x}) = \frac{K}{NV} \qquad\qquad K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

- Then

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right)$$

Where $\mathrm{V} = h^D$ denotes the volume of a hypercube of side h in D dimensions.

- Using the symmetry of the function $k(\mathbf{u})$, we can interpret this equation, not as a single cube centered on $\mathbf{x}$ but as the sum over N cubes centered on the N data points x_n.
- This kernel density estimator will suffer from the discontinuities at the boundaries of the cubes.
- We can obtain a smoother density model if we choose a smoother kernel function

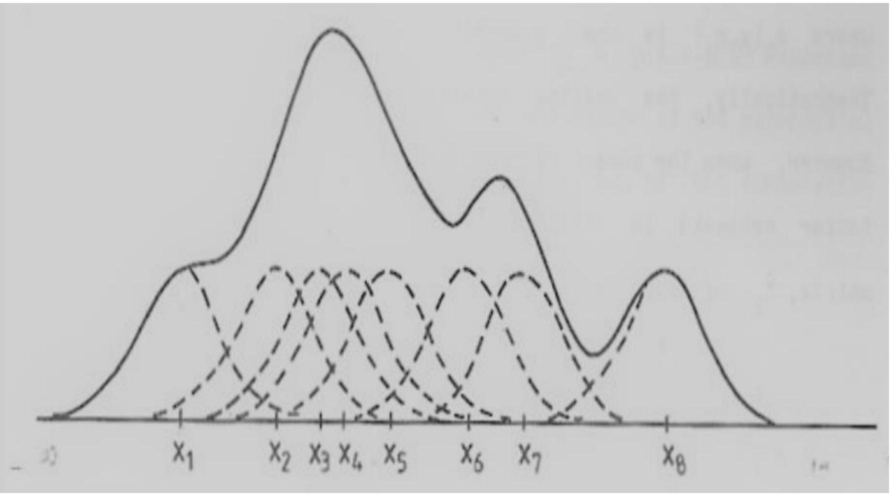
Kernel Density Estimators

- Common Choice: the Gaussian kernel function

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right)$$

  Where h now denotes the standard deviation of Gaussian components.

- This density model is obtained by placing a Gaussian over each data point and then adding up the contributions over the whole data set, and then dividing by N so that the density is correctly normalized.



Kernel Density Estimators

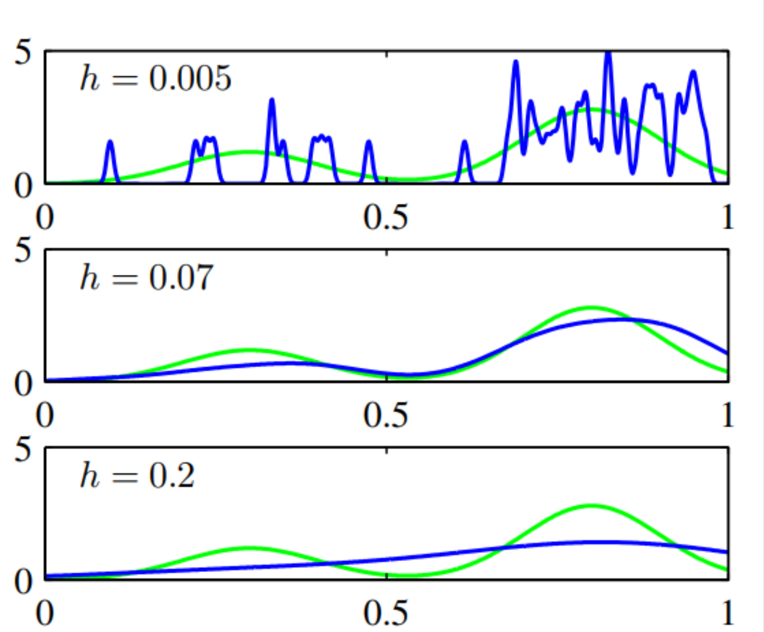


Figure: Three examples of density estimation corresponding to three different choices of h

- Data (50 observations) is drawn from a mixture of two Gaussians (Green curve)
- Small h, noisy density model with structure not in the distribution
- Large Δ, smooth density model without underlying bi-modality
- Best results from intermediate Δ

- We can choose any other kernel function $k(\mathbf{u})$ subject to the conditions

$$k(\mathbf{u}) \geq 0,$$

$$\int k(\mathbf{u})d\mathbf{u} = 1$$