Non-parametric density estimation and classification methods

Milos Hauskrecht
milos@cs.pitt.edu
5329 Sennott Square

Nonparametric Density Estimation Methods

- **Parametric distribution models** are:
  - restricted to specific forms, which may not always be suitable;
  - Example: modelling a multimodal distribution with a single, unimodal model.
- **Nonparametric approaches:**
  - make few assumptions about the overall shape of the distribution being modelled.

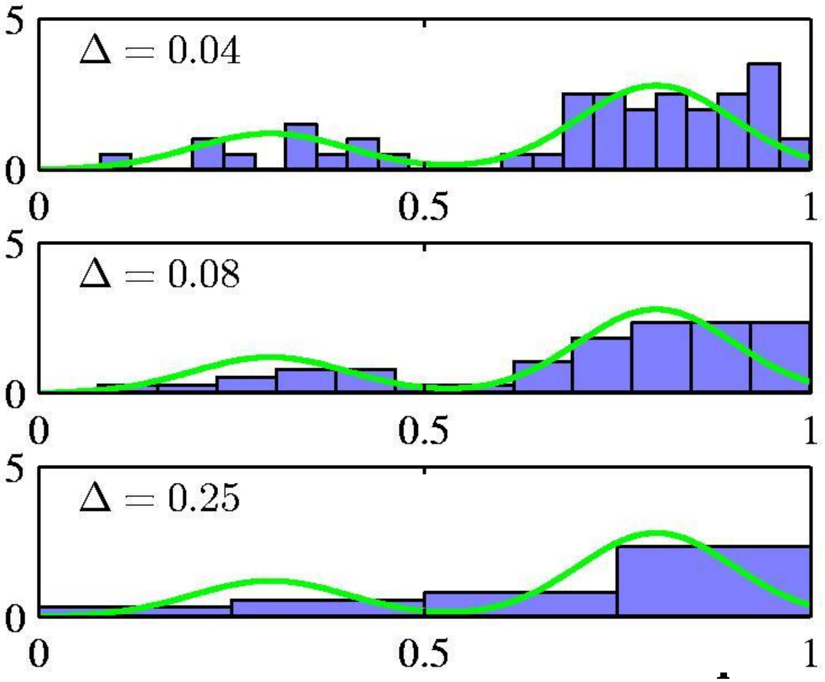
Nonparametric Methods

Histogram methods:

partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

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

- Often, the same width is used for all bins, $\Delta_i = \Delta$.
- Δ acts as a smoothing parameter.



M^D bins!

Nonparametric Methods

- Assume observations drawn from a density p(x) and consider a small region R containing x such that

$$P = \int_R p(x)dx$$

- The probability that K out of N observations lie inside R is *Bin(K,N,P)* and if N is large

$$K \cong NP$$

If the volume of R: denoted *V*, is sufficiently small, p(x) is approximately constant over R and

$$P \cong p(x)V$$

Thus

$$p(x) = \frac{P}{V}$$

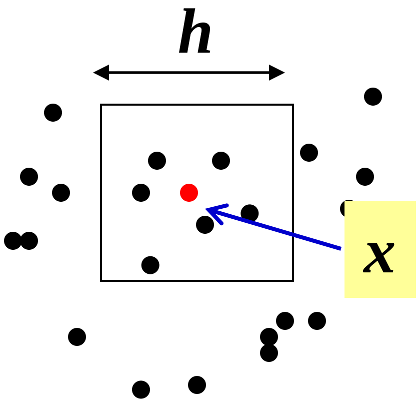
$$\boxed{p(x) = \frac{K}{NV}}$$

Nonparametric Methods: kernel methods

Kernel Density Estimation:

Fix V, estimate K from the data. Let R be a hypercube centred on **x** and define the kernel function (Parzen window)

$$k\left(\frac{x-x_n}{h}\right)= \begin{cases} 1 & |(x_i - x_{ni})|/h \le 1/2 \qquad i=1,\ldots D \\ 0 & otherwise \end{cases}$$



- **It follows that**

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

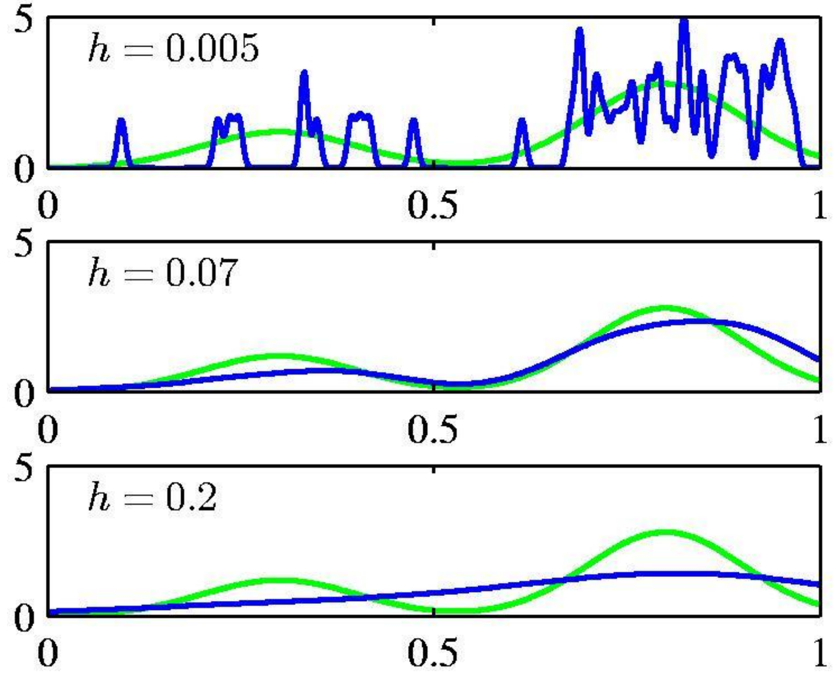
- **and hence**

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

Nonparametric Methods: smooth kernels

To avoid discontinuities in p(x) because of sharp boundaries use a **smooth kernel**, e.g. a Gaussian

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



h acts as a smoother.

- Any kernel such that

$$\begin{aligned} k(\mathbf{u}) &\geqslant 0, \\ \int k(\mathbf{u})\,\mathrm{d}\mathbf{u} &= 1 \end{aligned}$$

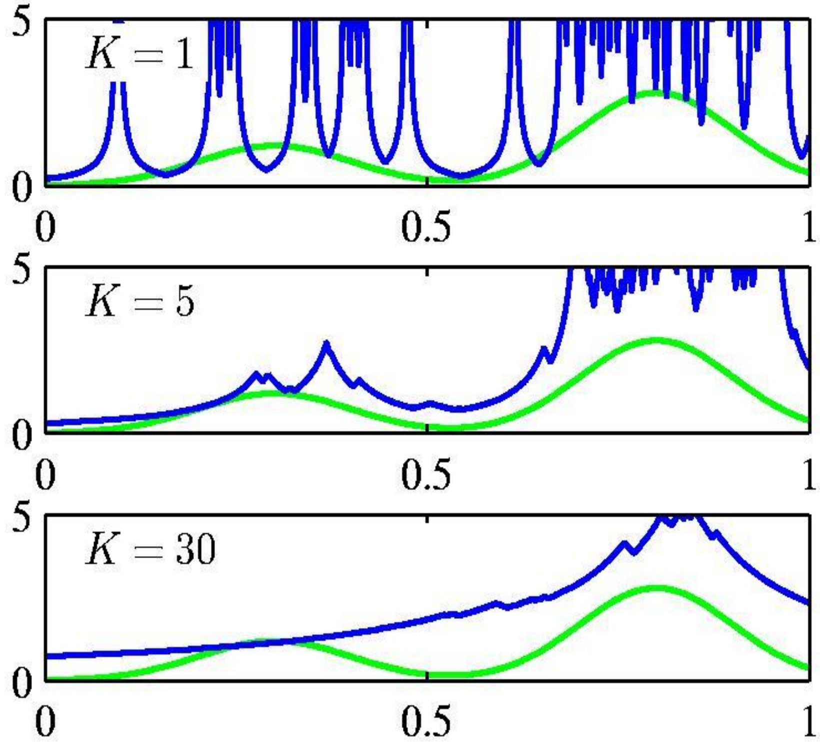
- will work.

Nonparametric Methods: kNN estimation

Nearest Neighbour Density Estimation:

fix K, estimate V from the data. Consider a hyper-sphere centred on x and let it grow to a volume, V*, that includes K of the given N data points. Then

$$p(\mathbf{x}) \simeq \frac{K}{NV^\star}.$$



K acts as a smoother

Nonparametric vs Parametric Methods

Nonparametric models:

- More flexibility – no density model is needed
- But require storing the entire dataset
- and the computation is performed with all data examples.

Parametric models:

- Once fitted, only parameters need to be stored
- They are much more efficient in terms of computation
- But the model needs to be picked in advance

Non-parametric Classification methods

- Given a data set with N_k data points from class C_k and $\sum_k N_k = N$, we have

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

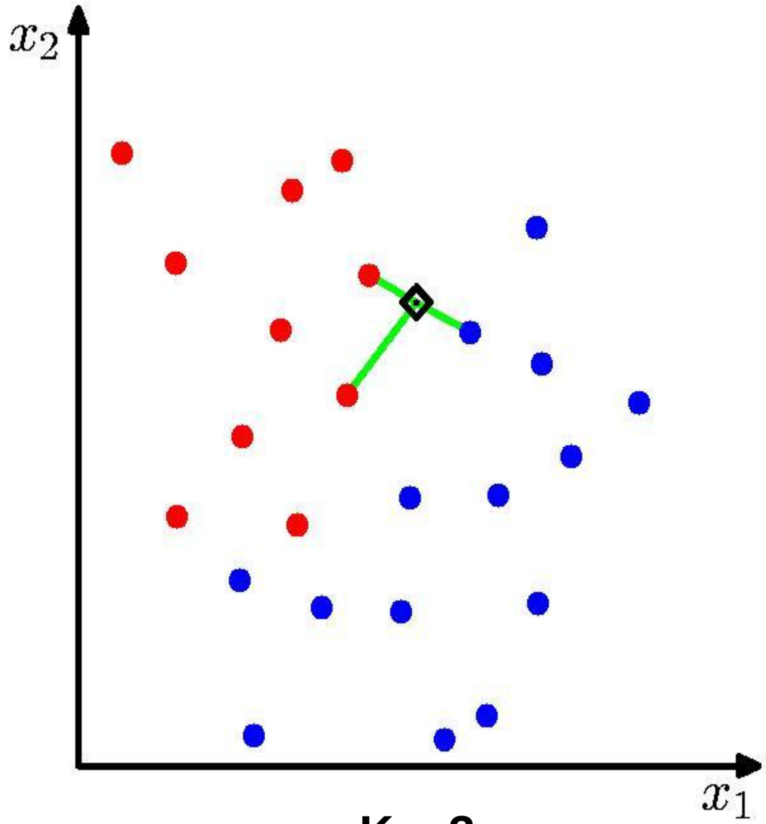
- and correspondingly

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{K_k}{N_k V}.$$

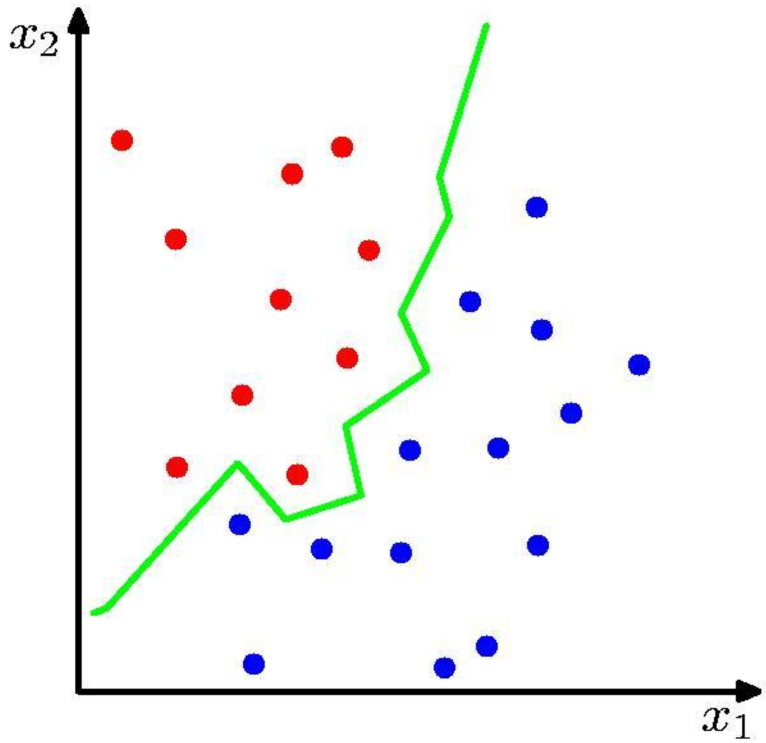
- Since $p(\mathcal{C}_k) = N_k/N$, Bayes' theorem gives

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{K_k}{K}.$$

K-Nearest-Neighbours for Classification



K = 3



Nonparametric kernel-based classification

- **Kernel function: k(x,x')**
  - Models similarity between x, x'
  - **Example:** Gaussian kernel we used in the kernel density estimation

$$k(x,x') = \frac{1}{(2\pi h^2)^{D/2}} \exp\left(-\frac{(x-x')^2}{2h^2} \right)$$

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} k(x, x_i)$$

- **Kernel for classification**

$$p(y = C_k \mid x) = \frac{\sum_{x':y'=C_k} k(x,x')}{\sum_{x'} k(x,x')}$$