CS 2750 Machine Learning
Lecture 6

Nonparametric density estimation

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

Parametric density estimation

Parametric density estimation:

- A set of random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$
- **A model of the distribution** over variables in ***X*** with **parameters** Θ : $\hat{p}(\mathbf{X} \mid \Theta)$
- **Data** $D = \{D_1, D_2, .., D_n\}$

Objective: find parameters Θ such that $p(\mathbf{X} \mid \Theta)$ describes data D the best

Parameter estimation (learning)

- **Maximum likelihood (ML)**

  $$\Theta_{ML} = \arg\max_{\Theta} p(D \mid \Theta, \xi)$$

- **Bayesian parameter estimation**

  keep the posterior density $p(\Theta \mid D, \xi)$

- **Maximum a posteriori probability (MAP)**

  $$\Theta_{MAP} = \arg\max_{\Theta} p(\Theta \mid D, \xi)$$

- **Expected value**

  $$\Theta_{EXP} = \int_{\Theta} \Theta p(\Theta \mid D, \xi) d\Theta$$

Nonparametric 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.

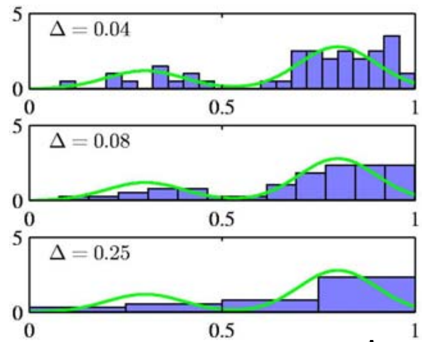
Nonparametric Methods

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, *V*, is sufficiently small, p(x) is approximately constant over R and

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

Thus

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

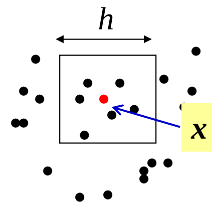
$$\boxed{p(x) = \frac{K}{NV}}$$

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)

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



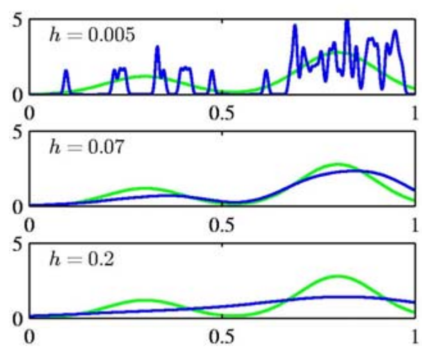
- **It follows that**
- **and hence** $K = \sum_{n=1}^{N} k\left(\frac{x - x_n}{h}\right)$

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

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

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

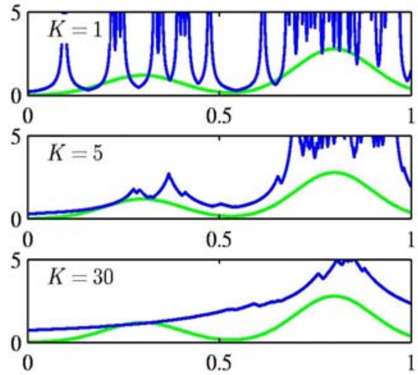
- will work.

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