NON-PARAMETRIC MODELING

J. Elder

CSE 6390/PSYC 6225 Computational Modeling of Visual Perception

Credits

- These slides were sourced and/or modified from:
  - Christopher Bishop, Microsoft UK

Nonparametric Methods

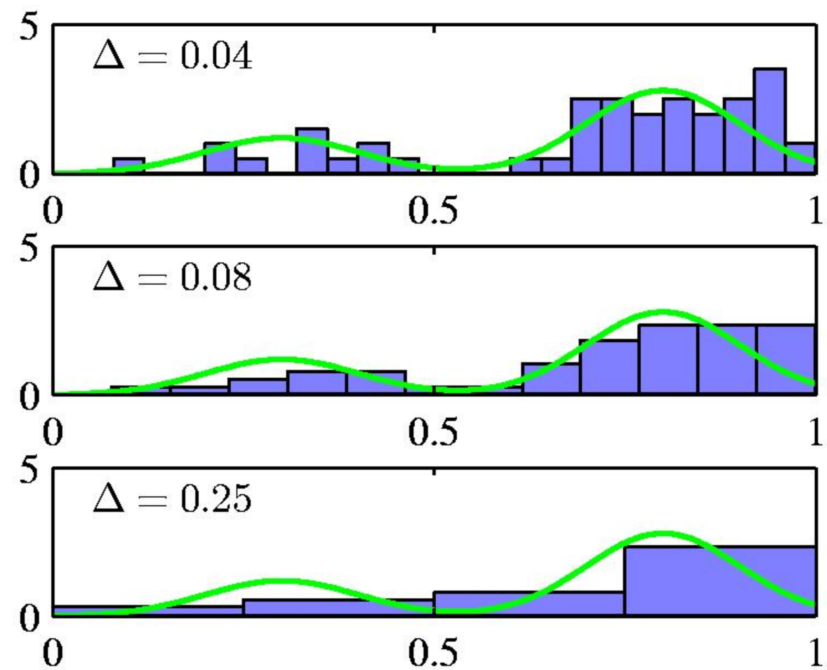
- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.
- Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

Histogramming

- **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.



- In a D-dimensional space, using M bins in each dimension will require M^D bins!

Kernel Density Estimation

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

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

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

$$K \simeq NP.$$

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

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

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

Kernel Density Estimation

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)

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

$$k((\mathbf{x}-\mathbf{x}_n)/h) = \begin{cases} 1, & |(x_i - x_{ni})/h| \leqslant 1/2, \qquad i = 1, \ldots, D, \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

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

and hence

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

Kernel Density Estimation

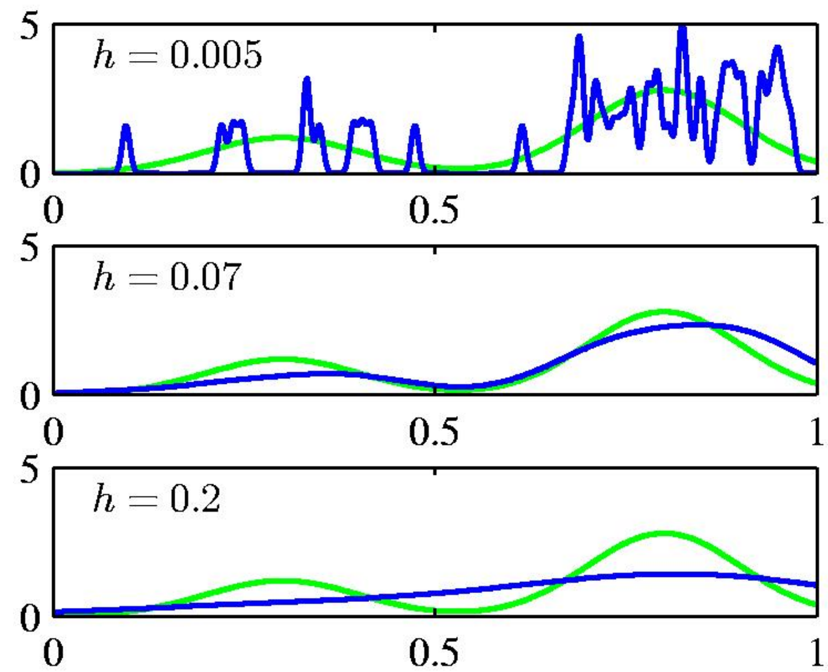
To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

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

(Any kernel such that

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

will work.)



h acts as a smoother.

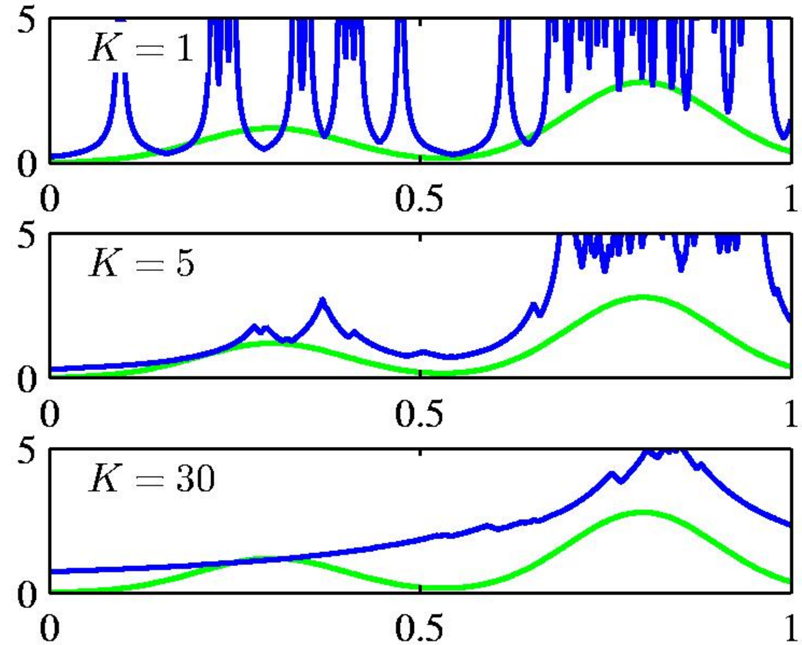
Kernel Density Estimation

- Problem: if V is fixed, there may be too few points in some regions to get an accurate estimate.

Nearest Neighbour Density Estimation

Nearest Neighbour Density Estimation: fix *K*, estimate *V* from the data. Consider a hypersphere 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.

Nearest Neighbour Density Estimation

- Problem: does not generate a proper density (for example, integral is unbounded on $\mathbb{R}^D$)
- In practice, on finite domains, can normalize.
- But makes strong assumption on tails $\left(\propto \frac{1}{x} \right)$

Nonparametric Methods

- Nonparametric models (not histograms) requires storing and computing with the entire data set.
- Parametric models, once fitted, are much more efficient in terms of storage and computation.

K-Nearest-Neighbours for Classification

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

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

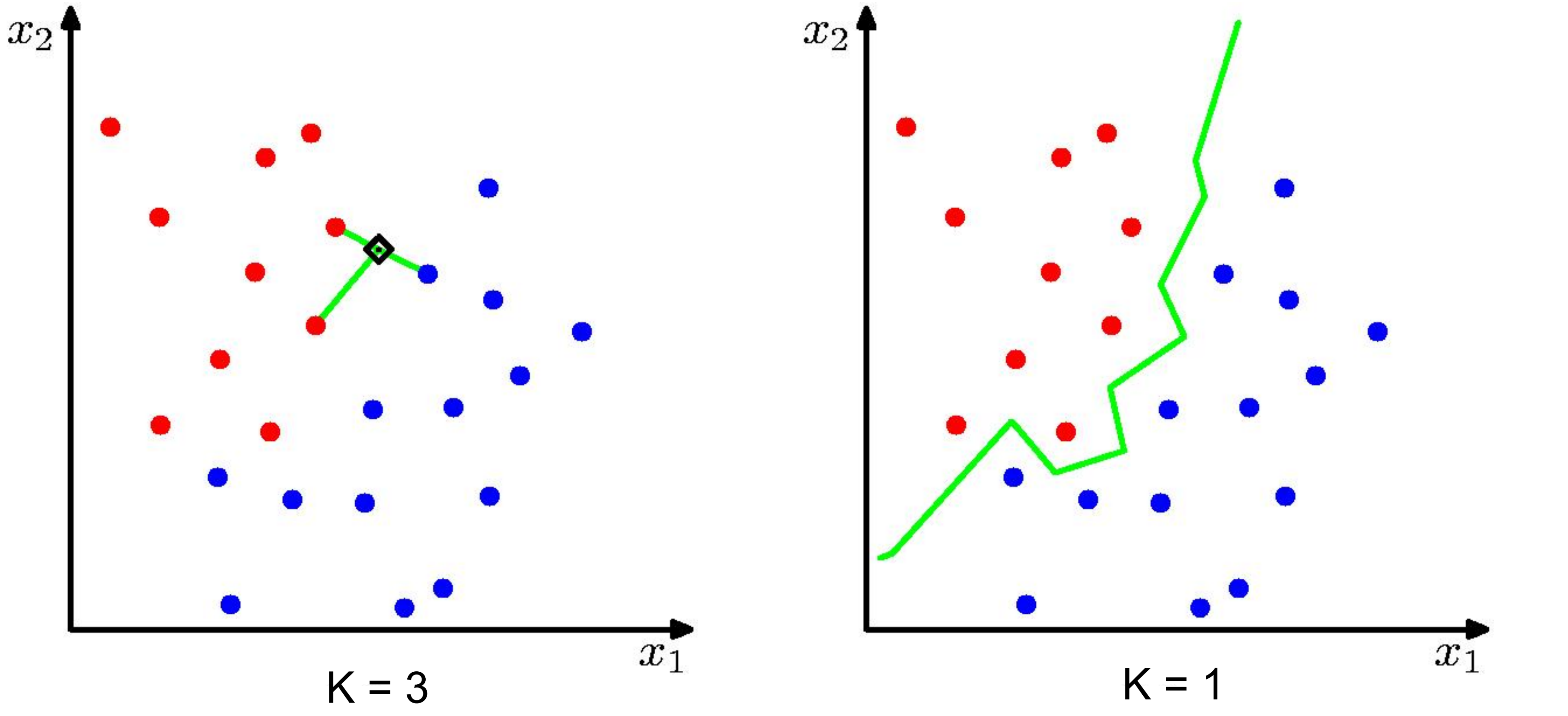
- and correspondingly

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

- 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-Nearest-Neighbours for Classification

- K acts as a smother
- As $N \to \infty$, the error rate of the 1-nearest-neighbour classifier is never more than twice the optimal error (obtained from the true conditional class