Advanced Data Analytics
Lecture week 3: Density Estimation

Ian T. Nabney

University of BRISTOL

Overview

- Basic understanding of nature and purpose of density estimation
- Two non-parametric density estimators: histograms and kernel estimators
- Strengths and weaknesses of non-parametric estimators

In the other lectures we will cover parametric estimation with mixture models and the EM algorithm, and latent variable methods including PCA (as a probabilistic model) and the Generative Topographic Mapping.

University of BRISTOL

Density estimation

- The task in density estimation is to model the probability distribution $p(\mathbf{x})$ of a random variable $\mathbf{x}$, given a finite set $\mathbf{x}_1, \ldots, \mathbf{x}_N$ of observations.
- We shall assume that the data points are independent and identically distributed.
- The problem of density estimation is fundamentally ill-posed, because there are infinitely many probability distributions that could have given rise to the observed finite data set. Indeed, any distribution $p(\mathbf{x})$ that is nonzero at each of the data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is a potential candidate. The issue of choosing an appropriate distribution relates to the problem of model selection.
- Most of the time we will consider parametric estimation, where we learn a small number of adaptive parameters.
- One limitation of the parametric approach is that it assumes a specific functional form for the distribution, which may turn out to be inappropriate for a particular application. An alternative approach is given by nonparametric density estimation methods in which the form of the distribution typically depends on the size of the data set. Such models still contain parameters, but these control the model complexity rather than the form of the distribution.

University of BRISTOL

Histogram methods

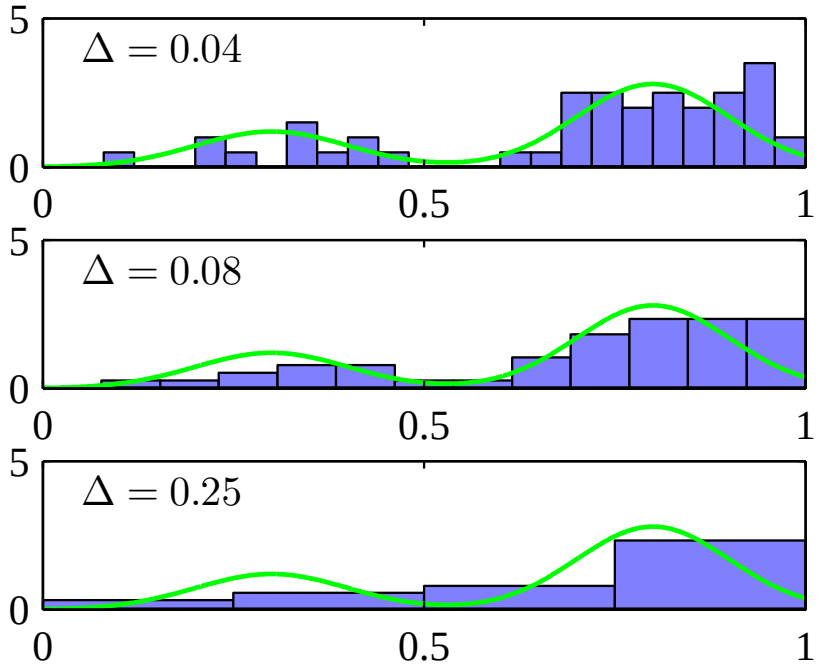
- We will consider a histogram as a density estimator for a single continuous variable x.
- Standard histograms simply partition x into distinct bins of width Δ_i and then count the number n_i of observations of x falling in bin i. In order to turn this count into a normalized probability density, we simply divide by the total number N of observations and by the width Δ_i of the bins to obtain probability values for each bin given by

$$p_i = \frac{n_i}{N\Delta_i} \tag{1}$$

- This gives a model for the density $p(x)$ that is constant over the width of each bin, and often the bins are chosen to have the same width $\Delta_i = \Delta$.

Histogram density estimation example

- A data set of 50 data points is generated from the distribution shown by the green curve.
- Histogram density estimates, based on (1), with a common bin width Δ are shown for various values of Δ.



Limitations of histogram estimation

- The histogram technique can be useful for obtaining a quick visualization of data in one or two dimensions but is unsuited to most density estimation applications.
- One obvious problem is that the estimated density has discontinuities that are due to the bin edges rather than any property of the underlying distribution that generated the data.
- Another major limitation of the histogram approach is its scaling with dimensionality. If we divide each variable in a D-dimensional space into M bins, then the total number of bins will be M^D. This exponential scaling with D is an example of the curse of dimensionality.
- In a space of high dimensionality, the quantity of data needed to provide meaningful estimates of local probability density would be prohibitive.

University of BRISTOL

Lessons for density estimation

- To estimate the probability density at a particular location, we should consider the data points that lie within some local neighbourhood of that point. This requires some form of distance measure, and here we have been assuming Euclidean distance (compare with t-SNE).
- For histograms, this neighbourhood property was defined by the bins, and there is a natural smoothing parameter describing the spatial extent of the local region, in this case the bin width. The value of the smoothing parameter should be neither too large nor too small in order to obtain good results.
- This is reminiscent of the choice of model complexity in polynomial curve fitting discussed in week 7 where the degree M of the polynomial, or the value λ of the regularization parameter, was optimal for some intermediate value, neither too large nor too small.

Kernel density estimation

- Start with a simple example of a small hypercube centred on each data point. In order to count the number K of points falling within this region, it is convenient to define the following function

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

  which represents a unit cube centred on the origin.

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

- The total number of data points lying inside a cube of side h centred on $\mathbf{x}$ is

$$K = \sum_{n=1}^{N} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right). \tag{3}$$

- The estimated density at $\mathbf{x}$

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{h^D} k\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \tag{4}$$

  where we have used $V = h^D$ for the volume of the hypercube.

University of BRISTOL

Gaussian kernel density estimator

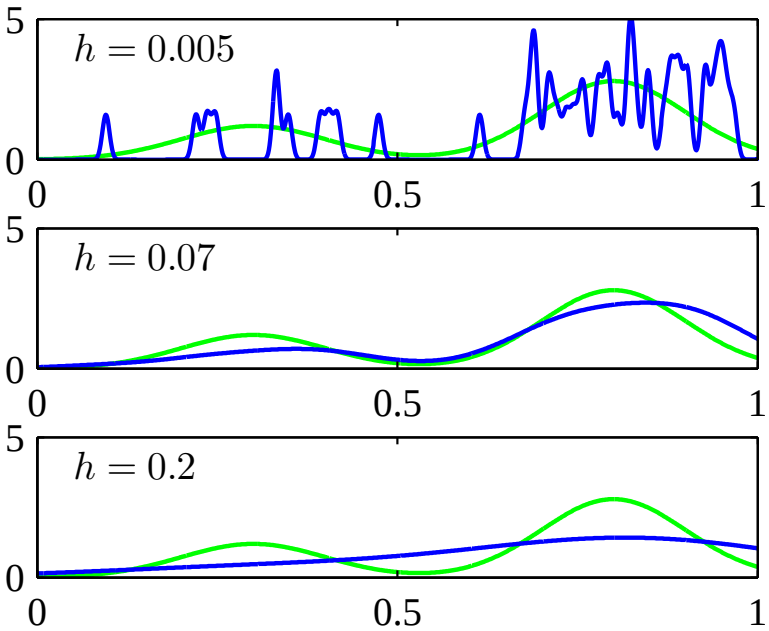
- This estimator also suffers from the presence of artificial discontinuities at the boundaries of the cubes.
- We can obtain a smoother density model if we choose a smoother kernel function

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{1/2}} \exp\left\{ -\frac{\|\mathbf{x} - \mathbf{x}_n\|^2}{2h^2} \right\} \tag{5}$$

  where h represents the standard deviation of the Gaussian components.
- Thus our 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 estimation example

- We see that h acts as a smoothing parameter and that if it is set too small (top panel), the result is a very noisy density model.
- If h is too large (bottom panel), then the bimodal nature of the underlying distribution from which the data is generated (shown by the green curve) is washed out.
- The best density model is obtained for some intermediate value of h (middle panel).



General kernel density estimation

- We can choose any other kernel function $k(\mathbf{u})$ in (4) subject to the conditions

$$k(\mathbf{u}) \geqslant 0, \tag{6}$$

$$\int k(\mathbf{u})\,\mathrm{d}\mathbf{u} = 1 \tag{7}$$

  which ensure that the resulting probability distribution is nonnegative everywhere and integrates to one.
- This method has the merit that there is no computation involved in the 'training' phase because this simply requires storage of the training set.
- However, this is also one of its great weaknesses because the computational cost of evaluating the density grows linearly with the size of the data set.
- Simple parametric models are very restricted in terms of the forms of distribution that they can represent. We therefore need to find density models that are very flexible and yet for which the complexity of the models can be controlled independently of the size of the training set: that is the mixture model.

University of BRISTOL

Density estimation for visualisation

- Bradypus variegatus, the Brown-throated Sloth.
- "Microryzomys minutus" , also known as the Forest Small Rice Rat, a rodent that lives in Peru, Colombia, Ecuador, Peru, and Venezuela.
- Note how the use of a smooth kernel provides a more realistic distribution map.