Non-parametric density estimation

Probability distributions

Francesco Corona

Non-parametric density estimation

We discussed probability distributions having specific functional forms governed by a number of parameters, whose values are to be determined from a data set

- This is called the parametric approach to density modelling

Limitation: The chosen density might be a poor model of the distribution that generates the data, which can result in poor predictive performance

- if the data generating process is multimodal, then this aspect of the distribution can never be captured by a Gaussian, which is unimodal

We consider some non-parametric approaches to density estimation that make few assumptions about the form of the distribution

- Focus mainly on simple frequentist methods

Outline

Histograms

Kernel density estimators

Nearest-neighbour methods
- Classification with k-NN

Histograms
Non-parametric density estimation

Histograms

Let us start with a discussion of histogram methods for density estimation

- Already encountered in the context of marginal/conditional distributions
- We explore the properties of histogram density models in some detail
- Focus on a single continuous variable x

Standard histograms simply partition x into distinct bins of width Δ_i

- then count the number n_i of observations of x falling in bin i

To turn this count into a normalised probability density, we divide n_i by the total number N of observations and by the width Δ_i of the bins

- We get probabilities values for each bin

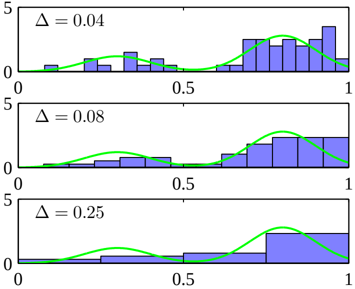
$$p_i = \frac{n_i}{N\Delta_i}, \qquad \text{such that} \int p(x)dx = 1 \tag{1}$$

This procedure gives a model for the density $p(x)$ that is constant over the bin

- The bins are often chosen to have the same width $\Delta_i = \Delta$

Histograms (cont.)

Data (50 observations) is drawn from the distribution, corresponding to the green curve, which is formed from a mixture of two Gaussians



Three examples of density estimates corresponding to three different choices of the bin width

- Small Δ, spiky density model with structure not in the distribution
- Large Δ, smooth density model without underlying bi-modality

Best results from an intermediate Δ

Useful technique for obtaining a quick visualisation of the data in 1 or $2D$

- Discontinuities, D variables divided in M bins each means M^D bins

Histograms (cont.)

Hardly useful in density estimation applications, but teaches important lessons

- To estimate the probability density at a particular location, we should consider data points that lie within some local neighbourhood of that point

The concept of locality needs some form of distance measure (Euclides does it)

- For histograms, locality was defined by the bins' width
- Locality should be neither too large nor too small

Kernel density estimation
Non-parametric density estimation

Kernel density estimators

Suppose our observations have been drawn from some unknown probability density $p(\mathbf{x})$ in some D-dimensional space, which we consider Euclidean

- We wish to estimate the value of $p(\mathbf{x})$

Let us consider some small region $\mathcal{R}$ containing $\mathbf{x}$

- The probability mass associated with this region is

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x} \qquad (2)$$

Suppose also that we have collected a dataset with N observations from $p(\mathbf{x})$

- Each point has a probability P of falling within $\mathcal{R}$

The total number of points K in R is distributed with a binomial distribution

$$\text{Bin}(K|N, P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{1-K} \qquad (3)$$

Kernel density estimators (cont.)

Using results for binomial distribution, the mean fraction of points in the region is $\mathbb{E}[K/N] = P$ and the variance around this mean is $\mathrm{var}[K/N] = P(1 - P)/N$

- For large N, the distribution will be sharply peaked arouns its mean

$$K \simeq NP \tag{4}$$

If we also assume that the region $\mathcal{R}$ is sufficiently small (of volume V) that the probability density is roughly constant over the region, then we have

$$P \simeq p(\mathbf{x})V \tag{5}$$

Combining the results, we obtain our density estimate in the form

$$p(\mathbf{x}) = \frac{K}{NV} \tag{6}$$

Kernel density estimators (cont.)

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

Either

- We can fix K and determine the value of V from the data
- We get the K-nearest-neighbour technique

or

- We can fix V and determine the value if K from the data
- We get a class of kernel-based techniques

In the limit $N \rightarrow \infty$, both techniques converge to the true probability density

- Provided that V shrinks suitably with N and that K grows with N

Kernel density estimators (cont.)

To start with we take the region $\mathcal{R}$ to be a small hypercube centred 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

$$k(\mathbf{u}) = \begin{cases} 1, & \text{if } |u_i| \leq 1/2 \quad \text{with } i = 1, \dots, D \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

It represents a unit cube centred on the origin

- Function $k(\mathbf{u})$ is an example of a **kernel function**
- In this context it is also called a **Parzen window**

If a data point x_n lies inside a cube of side h centred on $\mathbf{x}$, then the quantity $k(\mathbf{x} - \mathbf{x}_n)/h$ will be one and zero otherwise

- The total number of points lying inside this cube will be

$$K = \sum_{n=1}^{N} k\Big(\frac{\mathbf{x} - \mathbf{x}_n}{h}\Big) \tag{8}$$

Kernel density estimators (cont.)

Substitute $K = \sum_{n=1}^{N} k\Big(\dfrac{\mathbf{x}-\mathbf{x}_n}{h}\Big)$ in $p(\mathbf{x}) = \dfrac{K}{NV}$, the estimated density at $\mathbf{x}$ is

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\Big(\frac{\mathbf{x}-\mathbf{x}_n}{h}\Big) \tag{9}$$

where $h^D = V$ denotes the volume of the hypercube of side h in D dimensions

We can interpret this equation, not a single cube centred on $\mathbf{x}$,
but as the sum over N cubes centred on the N data points $\mathbf{x}_n$

This kernel density estimator has some of the same problems of the histograms

- Discontinuities, at the boundaries of the cubes

A smoother density model is obtained by choosing a smoother kernel function

Kernel density estimators (cont.)

Common choice: The kernel function of the density estimator is the Gaussian

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{n} \frac{1}{(2\pi h^2)^{1/2}} \exp\Big(- \frac{||\mathbf{x} - \mathbf{x}_n||^2}{2h^2}\Big) \tag{10}$$

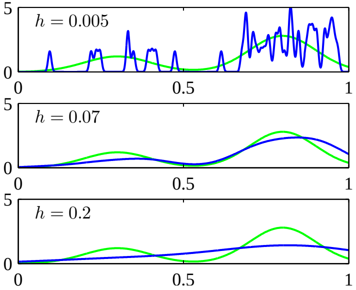
where h now denotes the standard deviation of the Gaussian components

This density model is obtained by placing a Gaussian over each data point, and then adding up the contributions over the whole dataset

- Divide by N to correctly normalise the density

Kernel density estimators (cont.)

Kernel density model applied to the same data set used with histograms



Three examples of density estimates:
3 different choices of h

- Small h, noisy density model with structure not in the distribution
- Large h, smooth density model without underlying bi-modality

Best results from an intermediate h

Parameter h plays the role of a smoothing term, and there is a trade-off between sensitivity to noise at small h and over-smoothing at large h

Kernel density estimators (cont.)

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

$$k(\mathbf{u}) \geq 0 \tag{11}$$

$$\int k(\mathbf{u}) d\mathbf{u} = 1 \tag{12}$$

They ensure that the resulting probability distribution is nonnegative everywhere and that integrates to one

Nearest-neighbour methods
Non-parametric density estimation

Nearest-neighbour methods

One of the difficulties with the kernel approach to density estimation is that the parameter h governing the kernel width is fixed for all kernels

- In regions of high density, a large h may lead to over-smoothing
- Reducing h, may lead to noisy estimates where the density is smaller

An optimal choice of h maybe dependent on the location within the data space

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

Instead of fixing V and determining the value of K from data, we consider a fixed value of K and use the data to find an appropriate value for V

Nearest-neighbour methods (cont.)

We consider a small sphere centred on point $\mathbf{x}$ at which we wish to estimate the density $p(\mathbf{x})$ and we allow the sphere to grow until it contains K data points

The estimate of the density is then given still by

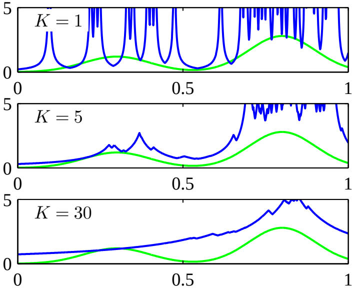
$$p(\mathbf{x}) = \frac{K}{NV}$$

with V set to the volume of the resulting sphere

This technique is known as **K-nearest neighbours**

The value of K now governs the degree of smoothing and there is an optimum choice for K that is neither too large nor too small

Nearest-neighbour methods (cont.)



The model produced by K-NN is not a true density model

- The integral over all space diverges ($\star$)

Classification with k-NN

The K-NN technique for density estimation can be used for classification

1. We apply it to each class separately
2. We make use of the Bayes' theorem

We have data: N_k points in class C_k with N total points such that $\sum_k N_k = N$

If we wish to classify a new point $\mathbf{x}$

Classification with k-NN (cont.)

1. Draw a sphere centred in $\mathbf{x}$ containing K points, whatever their class
2. Say, the volume of the sphere is V and contains K_k points of class C_k
3. Use $p(\mathbf{x}) = \dfrac{K}{NV}$ to estimate of the density associated with each class

$$p(\mathbf{x}|c_k) = \frac{K_k}{N_k V} \tag{13}$$

4. The unconditional density and the class prior are given by

$$\begin{aligned} p(\mathbf{x}) &= \frac{K}{NV} \qquad (14) \\ p(C_k) &= \frac{N_k}{N} \qquad (15) \end{aligned}$$

5. Combine Eq. 13, 14 and 15 using Bayes' theorem to get the posterior probability of the class membership

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K} \tag{16}$$

Classification with k-NN (cont.)

If we wish to minimise the probability of misclassification, we assign the test point $\mathbf{x}$ to the class having the largest posterior probability

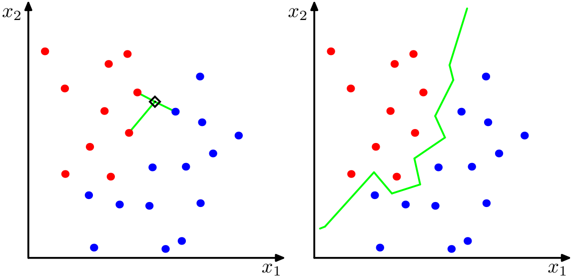
- The largest value of K_k/K

To classify $\mathbf{x}$, we identify the K nearest points from the training set and then assign it to the class with the largest number of representatives in this set

- Ties can be broken at random

Classification with k-NN (cont.)

In the K-NN classifier, a new point (black diamond), is classified according to the majority class membership of the K closest training points (here, $K = 3$)



(a) (b)

In the nearest-neighbour (K $=$ 1) approach to classification, the decision boundary is composed of hyperplanes that form perpendicular bisectors of pairs of points from different classes

Classification with k-NN (cont.)