Sampling and statistics

Basic inference

Francesco Corona

Department of Computer Science
Federal University of Ceará, Fortaleza

Sampling and statistics

Basic inference

Sampling and statistics

Concepts like samples and statistics seem to be all over the place

We introduce the main tools of orthodox inference

- ⇝ Confidence intervals
- ⇝ Hypothesis testing

Sampling and statistics (cont.)

Consider the typical statistical setting

There is a random variable X which we consider of interest

- Its PDF $f(x)$ or its PMF $p(x)$ are unknown

Roughly, our ignorance about $f(x)$ or $p(x)$ is one of two

- $f(x)$ or $p(x)$ is completely unknown
- The form of $f(x)$ or $p(x)$ is known

Let us consider first the second type of problem

- The form of $f(x)$ or $p(x)$ is known
- Down to a parameter θ

Sampling and statistics (cont.)

1. $X \sim \text{Exp}(\theta)$, θ is unknown
2. $X \sim \Gamma(\alpha, \beta)$, α and β are unknown
3. $X \sim b(n, p)$, n is known, p is known
4. $X \sim N(\mu, \sigma^2)$, μ and σ are unknown
5. ...

The RV X has a density or a mass function of the form $f(x|\theta)$ or $p(x|\theta)$

- $\theta \in \Omega$, for a specified set Ω

θ is the unknown parameter of the distribution

- We want to estimate it

Sampling and statistics (cont.)

Assume all information about the unknown distribution of X (or the unknown parameters of the distribution of X) comes from a **sample** on X

- The sample observations have the same (identical) distribution as X

We sample observations as the random variables $X_1, X_2, \dots, X_n$

- n indicates the **sample size**

When the sample is drawn, we use lower case letters $x_1, x_2, \dots, x_n$

- The values or **realisations** of the sample

Sampling and statistics (cont.)

Often, we can make reasonable assumptions about the sample observations

We can assume that $X_1, X_2, \ldots, X_n$ are also mutually independent RVs

⇝ In this case, we call the sample a **random sample**

Definition

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (IID) RVs

These random variables are said to constitute a ***random sample***

- *From the common distribution, and of size n*

■

Sampling and statistics (cont.)

Functions of the sample can be used to summarise the information in it

- Such sample functions are called **statistics**

Definition

Let $X_1, X_2, \ldots, X_n$ indicate a sample on a random variable X

Let $T = T(X_1, X_2, \ldots, X_n)$ be a function of the sample

Then, T is said to be a ***statistic***

When the sample is drawn, t is called a realisation of random variable T

$\rightsquigarrow$ $t = T(x_1, x_2, \ldots, x_n)$

($x_1, x_2, \ldots, x_n$ is a realisation of the sample)

Sampling and statistics (cont.)

Based on this terminology, we can formulate the problem we are developing

Let $X_1, X_2, \ldots, X_n$ denote a random sample on a RV X with density or mass function of the form $f(x)$ or $p(x)$, where $\theta \in \Omega$ for a specified set Ω

It makes some sense to consider a statistic T that is an ***estimator*** *of θ*

- *T is formally called a* ***point estimator*** *of θ*
- *Its realisation t is an* ***estimate*** *of θ*

Sampling and statistics (cont.)

Point estimators have several properties (we discuss some of them)

Definition

Unbiased-ness

Let $X_1, X_2, \dots, X_n$ *denote a sample on a RV* X *with PDF* $f(x|\theta)$, $\theta \in \Omega$

Let $T = T(x_1, x_2, \dots, x_n)$ *be a statistics*

We say that T *is an* ***unbiased estimator*** *of* θ *if* $E(T) = \theta$

Sampling and statistics (cont.)

We briefly discuss the **maximum likelihood estimator** (**MLE**)

- We start introducing the general concept of inference

We utilise the MLE to get point estimates for application problems

- We first discuss continuous case

Sampling and statistics (cont.)

Information in the sample and parameter θ are in the joint distribution

$$g(x_1, x_2, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta)$$

We can understand this symbol also as a function of θ

$$\mathcal{L}(\theta) = \mathcal{L}(\theta | x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i | \theta) \qquad (1)$$

This function is called the **likelihood function** of the random sample

Sampling and statistics (cont.)

As an estimate of θ, we might consider a measure of the centre of $\mathcal{L}(\theta)$

A common estimate is the value of θ that gives the maximum of $\mathcal{L}(\theta)$

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta) \tag{2}$$

This is the **maximum likelihood estimator** (**MLE**)

- If it is unique

Sampling and statistics (cont.)

Often it is convenient to work with the logarithm of the likelihood function

$$l(\theta) = \log \big[\mathcal{L}(\theta)\big]$$

The value of θ that maximises $l(\theta)$ is the same as the one that does $\mathcal{L}(\theta)$

- (As the log is a strictly increasing function)

Sampling and statistics (cont.)

For most of our models, the PDF/PMF is a differentiable function of θ

Thus, $\hat{\theta}$ frequently solves the equation

$$\frac{\partial l(\theta)}{\partial \theta} = 0 \tag{3}$$

Remark

For $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)'$ a vector of parameters, this is a system of equations

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1} &= 0 \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_2} &= 0 \\ \cdots &= 0 \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_d} &= 0 \end{aligned}$$

- They must be be solved simultaneously: $\nabla l(\boldsymbol{\theta}) = \mathbf{0}$

Sampling and statistics (cont.)

Under general conditions, MLEs are known to exhibit good properties

Suppose that we are not only interested in the parameter θ

- Say, we are also interested in parameter $\eta = g(\theta)$
- For some specified function g

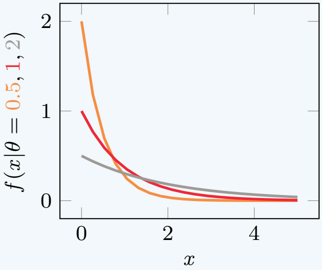
Then, the MLE of η is $\hat{\eta} = g(\hat{\theta})$, with $\hat{\theta}$ the MLE of θ

Sampling and statistics (cont.)

Example

Exponential distribution

The common distribution of random sample $X_1, X_2, \ldots, X_n$ is the $\Gamma(1, \theta)$



$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x \in (0, \infty)$$

Sampling and statistics (cont.)

The log of the likelihood function

$$l(\theta) = \log\Big(\prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta}\Big) = -n\log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

The first partial derivative of the log-likelihood with respect to θ

$$\frac{\partial l(\theta)}{\partial \theta} = -n\theta^{-1} + \theta^{-2}\sum_{i=1}^{n} x_i$$

Setting the derivative to 0 and solving for θ, we obtain the solution $\overline{x}$

Sampling and statistics (cont.)

There is only one critical value

The second partial of the log-likelihood at $\overline{x}$ is strictly negative

- This verifies that $\overline{x}$ gives a maximum

Hence, $\hat{\theta} = \overline{X}$ is the MLE of θ

Because $E(X) = \theta$, we have that $E(\overline{X}) = \theta$

$\rightsquigarrow$ $\hat{\theta}$ is an unbiased estimator of θ

■

Sampling and statistics (cont.)

Example

Binomial distribution

Let X be 1 or 0, depending on the outcome of a Bernoulli experiment

Let θ with $0 < \theta < 1$ indicate the probability of success

The PMF of X

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}, \quad x = 0 \text{ or } 1$$

If $X_1, X_2, \ldots, X_n$ is a random sample on X, then the likelihood function

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}, \quad x_i = 0 \text{ or } 1$$

Sampling and statistics (cont.)

Taking logarithms, we get

$$l(\theta) = \sum_{i=1}^{n} x_i \log(\theta) + \Big(n - \sum_{i=1}^{n} x_i\Big) \log(1-\theta), \quad x_i = 0 \text{ or } 1$$

The partial derivative of $l(\theta)$

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta}$$

Setting it to 0 and solving for θ

$$\hat{\theta} = n^{-1} \sum_{i=1}^{n} X_i = \overline{X}$$

The MLE is the proportion of successes in the n trials

Because $E(X) = \theta$, we have that $E(\overline{X}) = \theta$

- $\hat{\theta}$ is an unbiased estimator of θ

Sampling and statistics (cont.)

Example

Normal distribution

Let X have a $N(\mu, \sigma^2)$ distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big[-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)\Big], \quad x \in (-\infty, +\infty)$$

In this case, $\boldsymbol{\theta} = (\mu, \sigma)'$

If $X_1, X_2, \ldots, X_n$ is a random sample on X, the log-likelihood function

$$l(\mu, \sigma) = -\frac{n}{2} \log{(2\pi)} - n \log{(\sigma)} - \frac{1}{2} \sum_{i=1}^{n} \Big(\frac{x_i - \mu}{\sigma}\Big)^2 \tag{4}$$

Sampling and statistics (cont.)

The two partial derivatives

$$\begin{aligned}\frac{\partial l(\mu,\theta)}{\partial \mu} &= -\sum_{i=1}^{n}\Big(\frac{x_i-\mu}{\sigma}\Big)\Big(-\frac{1}{\sigma}\Big)\\ \frac{\partial l(\mu,\theta)}{\partial \theta} &= -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned} \tag{5}$$

Setting them to zero and solving simultaneously, we get the MLEs

$$\begin{aligned}\hat{\mu} &= \overline{X}\\ \hat{\sigma}^2 &= n^{-1}\sum_{n=1}^{n}(X_i-\overline{X})^2\end{aligned} \tag{6}$$

Note that we used the fact that the MLE of σ^2 is the MLE of σ squared

Sampling and statistics (cont.)

$\hat{\mu}$ is an unbiased estimator of μ

$\hat{\sigma}_2$ is a biased estimator of σ^2

- The bias of $\hat{\sigma}^2$ is $E(\hat{\sigma}^2 - \sigma^2) = -\sigma^2/n$

It converges to zero as $n \to \infty$

Nonparametric density estimation

Sampling and statistics

Non-parametric density estimation

We only considered probability distributions with specific functional forms

- Functions governed by a number of parameters, to be estimated

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 the (unimodal) normal

Non-parametric density estimation

We consider some **non-parametric** approaches to density estimation

- Very few assumptions about the form of the distribution
- Focus mainly on simple frequentist methods

Histogram estimates

Non parametric densities

Histogram estimates

Let $X_1, X_2, \dots, X_n$ be a random sample on a RV X with CDF $F(x)$

We briefly discuss a histogram of the sample

- An estimate of the PMF/PDF of X

We do not make assumptions on the form of the distribution

- We only say whether they are discrete or continuous

The histogram is a **non-parametric estimator**

The distribution of X is discrete

Histogram estimates

The distribution of X is discrete (cont.)

Assume that X is a discrete random variable with the PMF $p(x)$

Suppose, first, that the range of X is finite

- $\mathcal{D} = \{a_1, \dots, a_m\}$

An informal estimator of $p(a_j)$ is the relative frequency of observations a_j

For $j = 1, 2, \dots, m$, we can define the statistics

$$I_j(X_i) = \begin{cases} 1, & X_i = a_j \\ 0, & X_i \neq a_j \end{cases}$$

The intuitive estimate of $p(a_j)$ is the average

$$\hat{p}(a_j) = \frac{1}{n} \sum_{i=1}^{n} I_j(X_i) \quad (7)$$

The distribution of X is discrete (cont.)

Estimates $\{\hat{p}(a_1), \hat{p}(a_2), \ldots, \hat{p}(a_m)\}$ are a nonparametric estimate of $p(x)$

- $I_j(X_i)$ has a Bernoulli distribution with probability of success $p(a_j)$

The distribution of X is discrete (cont.)

Suppose now that the space of X is infinite, $\mathcal{D} = \{a_1, a_2, \dots\}$

We select a value, say a_m, and we make the groupings

$$\{a_1\}, \{a_2\}, \dots, \{a_m\}, \tilde{a}_{m+1} = \{a_{m+1}, a_{m+2}, \dots\} \tag{8}$$

Let $\hat{p}(\tilde{a}_{m+1})$ be the proportion of sample observations that $\geq a_{m+1}$

Estimates $\{\hat{p}(a_1), \hat{p}(a_2), \dots, \hat{p}(a_{m+1}), \hat{p}(\tilde{a}_{m+1})\}$ form the estimate of $p(x)$

The distribution of X is discrete (cont.)

A rule of thumb for group merging

Select m so that the frequency of category a_m exceeds twice the combined frequencies of categories $a_{m+1}, a_{m+2}, \ldots$

The distribution of X is discrete (cont.)

A **histogram** is a barpolot of $\hat{p}(a_j)$ versus a_j

There are two cases two consider

1. The values a_j represent qualitative categories
2. The values of a_j represent ordinal information

The distribution of X is discrete (cont.)

Example

The hair of young brits

Five hair colour were recorded for a sample size $n = 50000$

	Fair	Red	Medium	Dark	Black
Count	12950	2950	21 500	12 700	350
$\hat{p}(a_j)$	0.259	0.059	0.421	0.254	0.007

The frequency distribution of this sample and the estimate of the PMF are

■

The distribution of X is discrete (cont.)

Example

Poisson variates

Consider 30 data that are simulated values drawn from discrete distribution

- A Poisson distribution with mean $\lambda = 2$

$$p(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{elsewhere} \end{cases}$$

2 1 1 1 1 5 1 1 3 0 2 1 1 3 4
2 1 2 2 6 5 2 3 4 1 3 1 3 0

The nonparametric estimate of the PMF

j	0	1	2	3	4	5	≥ 6
$\hat{p}(a_j)$	0.067	0.367	0.233	0.167	0.067	0.067	0.033

The distribution of X is continuous

Histogram estimates

The distribution of X is continuous

Assume the random sample $X_1, \dots, X_n$ from a continuous RV X, PDF $f(t)$

We firstly sketch an estimate for this PDF at some given value x

- Then, we use the estimate to develop a histogram of the PDF

For an arbitrary but fixed point x and a given $h > 0$, consider the interval

$$(x - h, x + h)$$

By the mean-value theorem for integrals, for some ξ with $|x - \xi| < h$,

$$P(x - h < X < x + h) = \int_{x+h}^{x-h} f(t) \mathrm{d}t = f(\xi) 2h \approx f(x) 2h$$

The distribution of X is continuous

$$P(x - h < X < x + h) = \int_{x+h}^{x-h} f(t)\mathrm{d}t = f(\xi)2h \approx f(x)2h$$

The nonparametric estimate of the LHS

It is the proportion of sample observations that fall in $(x - h, x + h)$

This suggests the nonparametric estimate of $f(x)$ at a given point x

$$\hat{f}(x) = \frac{1}{2h} \frac{\#\{x - h < X_i < x + h\}}{n} \qquad (9)$$

The distribution of X is continuous (cont.)

More formally, we consider the indicator statistic

$$I_i(x) = \begin{cases} 1, & x - h < X_i < x + h \\ 0, & \text{otherwise} \end{cases}, \quad i = 1, \dots, n$$

Then the nonparametric estimator of $f(x)$ becomes

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} I_i(x) \tag{10}$$

Since the sample observations are identically distributed

$$E[\hat{f}(x)] = \frac{1}{2hn} n f(\xi) 2h = f(\xi) \to f(x), \quad \text{as } h \to 0$$

Hence $\hat{f}(x)$ is approximately an unbiased estimator of the density $f(x)$

The distribution of X is continuous (cont.)

The indicator function I_i is called the **rectangular kernel**

- $2h$ is the **bandwidth**

The distribution of X is continuous (cont.)

Let $x_1, x_2, \dots, x_n$ be the realised values of the random sample

The histogram estimate of $f(x)$ is obtained as follows

Opposite to the discrete case, classes for the histogram must be selected

One way of doing this

- Select a positive integer m
- Select an $h > 0$
- Select a value a such that $a < \min(x_i)$

The m intervals below must cover the sample range $\big[\min(x_i), \max(x_i)\big]$

$$\big(a-h, a+h\big], \big(a+h, a+3h\big], \big(a+3h, a+5h\big], \cdots, \\ \big(a+(2m-3)h, a+(2m-1)h\big] \qquad (11)$$

These intervals form the histogram classes

The distribution of X is continuous (cont.)

For the histogram

Consider the i-th interval, $\big(a+(2i-3)h, a+(2i-1)h\big]$ with $i = 1, 2, \dots, m$

- Over the interval, let the height of the bar be the density estimate $\hat{f}(x)$

$$\hat{f}\big[a + 2(i-1)h\big]$$

  That is, at the mid-point of the interval

- The height of the bar is thus proportional to the number of x_is that fall in the interval $\big(a + (2i-3)h, a + (2i-1)h\big]$

To complete the histogram estimate of $f(x)$

- 0 for $x \leq a$
- 0 for $x > a + (2m-1)h$

The distribution of X is continuous (cont.)

Let I_i be the intervals of the partition

$$I_i = \big(a + (2i-3)h, a + (2i-1)h\big], \quad i = 1, \dots, m$$

Then, we can summarise the histogram estimate of the PDF

$$\hat{f} = \begin{cases} \#\big\{a + (2i-3)h < X_i \leq a + (2i-1)h\big\}/(2hn), & x \in I_i, i = 1, \dots, m \\ 0, & \text{elsewhere} \end{cases} \tag{12}$$

The estimator is non-negative and it integrates to one over $(-\infty, +\infty)$

- The properties of a PDF are satisfied

The distribution of X is continuous (cont.)

Histograms partition x into distinct bins of potentially different widths Δ_i

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

This count needs be turned into a normalised probability density

- We divide n_i by the total number N of observations and by the width Δ_i

We get the probabilities values for each of the bins

$$p_i = \frac{n_i}{N \Delta_i}, \qquad \text{such that} \int p(x) dx = 1 \tag{13}$$

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

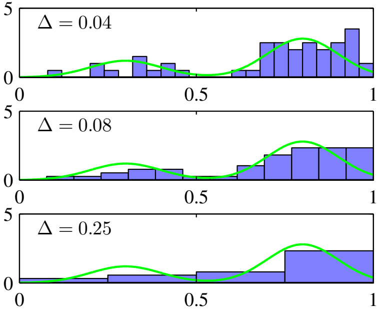
- The bins are often chosen to have the same width $\Delta_i = \Delta$

Histograms (cont.)

Data (50 observations) is drawn from some distribution (the green curve)

- A mixture of two normals

Three density estimates with three different choices of bin width Δ



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

Useful technique for getting 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 it teaches a lessons

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

The **notion of locality** needs some form of **distance measure**

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

Kernel density estimates

Non parametric densities

Kernel density estimators

Suppose our observations have been drawn from some unknown 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 associated with this region

$$P = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x} \tag{14}$$

Kernel density estimators (cont.)

Suppose that we have a random sample with N observations from $p(\mathbf{x})$

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

The number of points K in $\mathcal{R}$ is distributed with a binomial distribution

$$\text{Bin}(K|N,P) = \frac{N!}{K!(N-K)!}P^K(1-P)^{1-K} \tag{15}$$

⇝ The mean fraction of points in the region

$$E(K/N) = P$$

⇝ The variance around this mean

$$\text{Var}(K/N) = P(1-P)/N$$

Kernel density estimators (cont.)

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

$$K \simeq NP \tag{16}$$

Assume that the region $\mathcal{R}$ is sufficiently small (of volume V)

- The probability density is roughly constant over the region

$$P \simeq p(\mathbf{x})V \tag{17}$$

Combining results, we obtain a density estimate in the form

$$p(\mathbf{x}) = \frac{K}{NV} \tag{18}$$

Kernel density estimators (cont.)

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

Option 1

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

Option 2

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

For $N \to \infty$, both techniques converge to the true probability density

Kernel density estimators (cont.)

Suppose that we take the region $\mathcal{R}$ to be a small hypercube

- Centred on some point $\mathbf{x}$
- (where we wish the density)

To count the number K of points falling within $\mathcal{R}$, define the function

$$k(\mathbf{u}) = \begin{cases} 1, & \text{if } |u_i| \leq 1/2 \quad \text{with } i = 1, \dots, D \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

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**

Kernel density estimators (cont.)

Suppose that a data point $\mathbf{x}_n$ lies inside a cube of side h centred on $\mathbf{x}$

Then, the quantity $k\Big(\dfrac{\mathbf{x}-\mathbf{x}_n}{h}\Big)$ will be one and zero otherwise

The total number of points lying inside this cube

$$K = \sum_{n=1}^{N} k\Big(\frac{\mathbf{x}-\mathbf{x}_n}{h}\Big) \tag{20}$$

Kernel density estimators (cont.)

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

We obtain a estimate of the density at $\mathbf{x}$

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{h^D} k\Big(\frac{\mathbf{x}-\mathbf{x}_n}{h}\Big) \qquad (21)$$

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

We can interpret this equation

- Not a single cube centred on $\mathbf{x}$
- The sum over N cubes centred on the N data points $\mathbf{x}_n$

Kernel density estimators (cont.)

Remark

This density estimator shares some of the problems of the histograms

- Discontinuities, at the boundaries of the cubes

A smoother model is obtained by choosing a smoother kernel function

Kernel density estimators (cont.)

The kernel function of the estimator is often chosen to be the Gaussian

$$p(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{n} \frac{1}{(2\pi h^2)^{D/2}} \exp\Big(- \frac{||\mathbf{x} - \mathbf{x}_n||^2}{2h^2} \Big) \qquad (22)$$

h denotes the standard deviation of Gaussian components

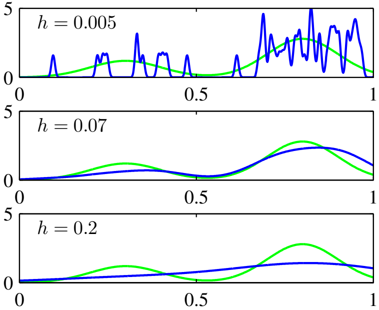
This density model is obtained by placing a Gaussian over each data point

- Then, adding up the contributions over the whole dataset
- And, dividing by N to correctly normalise the density

Kernel density estimators (cont.)

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

Three density estimates with three different choices of h



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

Parameter h plays the role of a smoothing term

- There is a trade-off
- 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

$$k(\mathbf{u}) \geq 0 \tag{23}$$

$$\int k(\mathbf{u}) d\mathbf{u} = 1 \tag{24}$$

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

Nearest neighbours methods

Non parametric densities

Nearest-neighbour methods

One of the difficulties with the kernel approach to density estimation

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 density is low

An optimal choice of h may be dependent on location within the space

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

We consider a fixed value of K and use the data to find a value for V

- Instead of fixing V and determining K from data

Nearest-neighbour methods (cont.)

Let $\mathcal{B}(\mathbf{x})$ be a ball centred on point $\mathbf{x}$ at which we wish to estimate $p(\mathbf{x})$

- Let the ball grow until it contains K points

The **K-nearest neighbours** density estimate

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

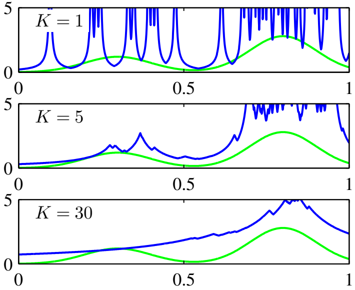
V is the volume of the resulting ball

There is an optimum choice for the value of K

- Neither too large nor too small

Nearest-neighbour methods (cont.)

The value of K governs the degree of smoothing of the estimate



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

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

Nearest-neighbour methods (cont.)

Example

The K-NN density estimator can be used for classification

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

We got 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 $\mathbf{x}$

Nearest-neighbour methods (cont.)

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

$$p(\mathbf{x}|c_k) = \frac{K_k}{N_k V} \tag{25}$$

4. The unconditional density and the class prior

$$\begin{aligned} p(\mathbf{x}) &= \frac{K}{NV} \\ p(C_k) &= \frac{N_k}{N} \end{aligned} \tag{26}$$

5. Combine the equations above using Bayes' theorem rule

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} = \frac{K_k}{K} \tag{27}$$

This is the posterior probability of the class membership

Nearest-neighbour methods (cont.)

If we wish to minimise the probability of misclassification

We assign the query point $\mathbf{x}$ to the class with largest posterior probability

- The largest value of K_k/K

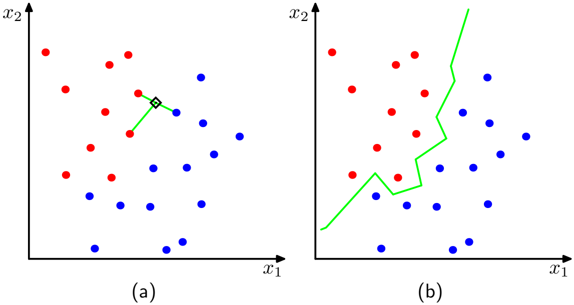
To classify $\mathbf{x}$, we identify the K nearest points from the training set

We assign it to the class with largest number of representatives in this set

- Ties can be broken at random

Nearest-neighbour methods (cont.)

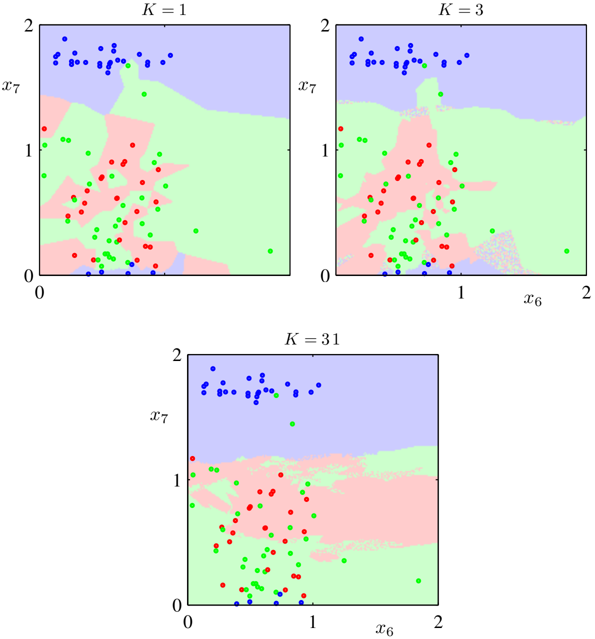
In the K-NN classifier, a new point (black), is classified according to the majority class membership of the K closest training points (here, $K = 3$)



The nearest-neighbour ($K = 1$) approach to classification

- The decision boundary is composed of hyperplanes

They form perpendicular bisectors of pairs of points from different classes


$K = 1$
$K = 3$
$K = 31$
x_7
x_6
0
1
2