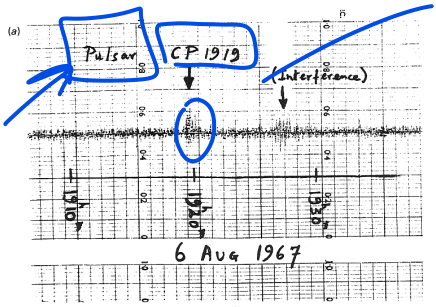
Queen Mary School Hainan
Queen Mary University of London

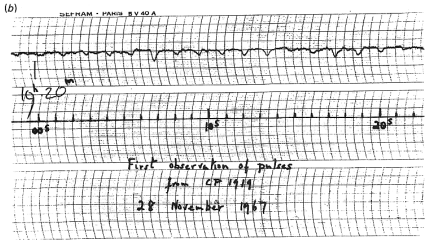
QHM5703 Principles of Machine Learning
Unsupervised learning: Density estimation

Dr Nikesh Bajaj
Dr Jesús Requena Carrión

Week 15: 15/16 Dec 2025

Bell's bit of scruff



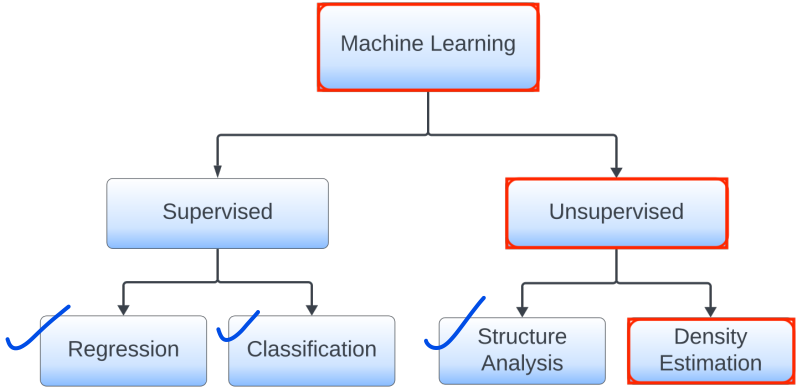


LGM

"I logged it with a question mark and moved on"
... Jocelyn Bell

Re-examine your data!

Machine Learning taxonomy



Outlier

Agenda

$P(x=1) = 3/4$

$p(x)$

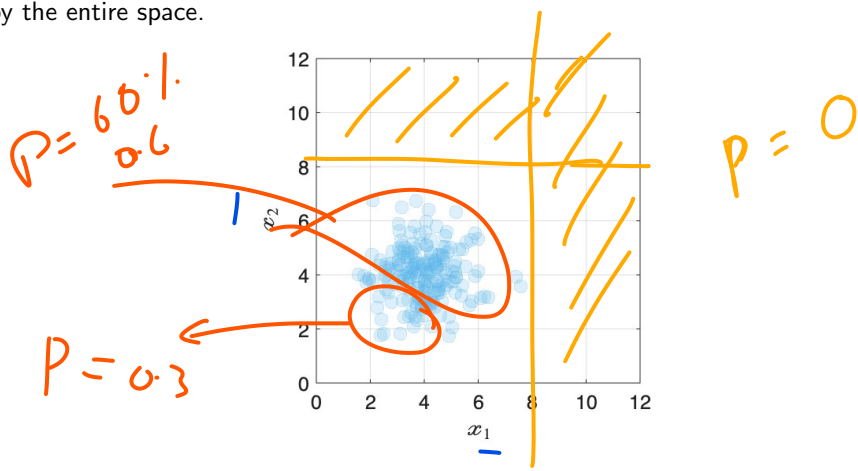
$\int_{-\infty}^{\infty} p(x)dx = 1$

1

$x-\delta \leq x \leq x+\delta$

Data distribution

Our datasets are collection of samples **distributed** in the attribute space. However, they do not occupy the entire space.



Data distribution

The basic observation that samples are not to be found anywhere in the attribute space, leads us to questions such as:

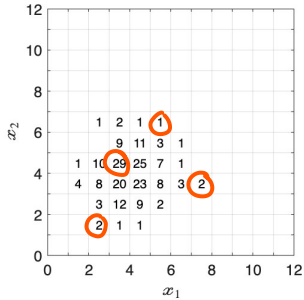
- Where are most of my samples?
- Should I expect a sample in this region of the attribute space?
- What is the probability that I will find a sample in this region?
- Given a probability, where will I find my next sample?

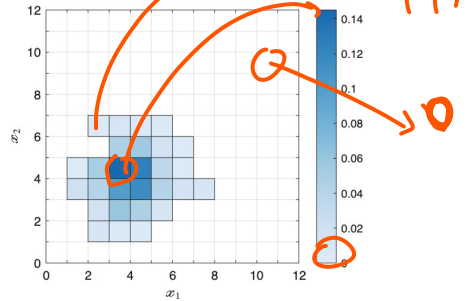
And most importantly, is there a **single approach** to answer them?

Divide and count

Partitioning the space into smaller regions and **counting** the number of samples within each region is a simple way of describing the **observed distribution**.

It also gives an indication of what to expect if we extract more samples from the same population, i.e. the **true distribution**. Counts will not give useful **quantitative** answers, but we can transform them into **rates**.

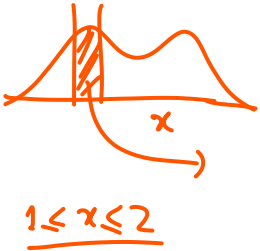




Probability density

Probability densities are models that describe the underlying **true distribution** of our data. They allow us to quantify:

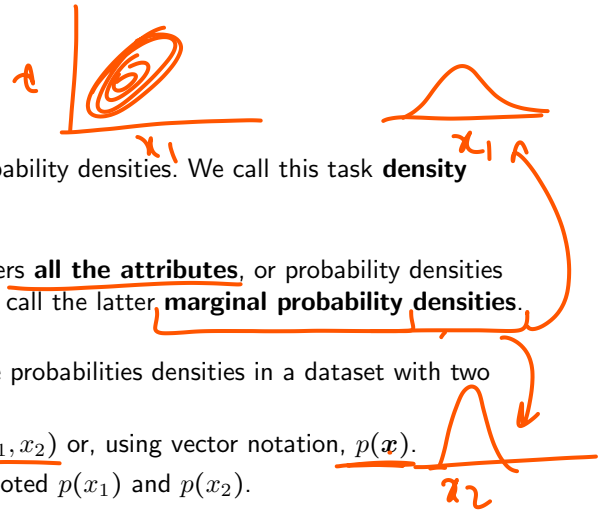
- The probability of extracting a sample within a region of the space.
- The fraction of population samples that lie within a region of the space.



region → Model → probability

The probability of finding a sample anywhere in the attribute space is 1. This probability is not uniform: it is **denser** in some regions than others.

Density estimation



In machine learning we **use data** to build probability densities. We call this task **density estimation**.

We can build a probability density that considers **all the attributes**, or probability densities that consider a **subset of the attributes**. We call the latter **marginal probability densities**.

Our (simplified) mathematical notation for the probabilities densities in a dataset with two attributes x_1 and x_2 will be:

- The probability density is denoted by $p(x_1, x_2)$ or, using vector notation, $p(\boldsymbol{x})$.
- The marginal probability densities are denoted $p(x_1)$ and $p(x_2)$.

This notation can be easily extended to larger number of attributes.

Agenda

Probability densities

Non-parametric methods

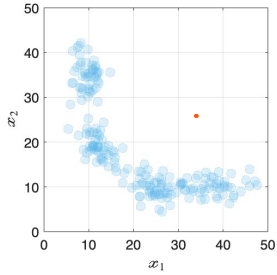
Parametric density estimation

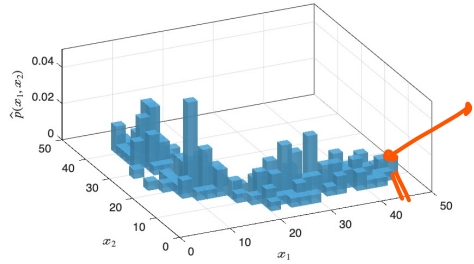
Applications

The histogram

Non-parametric methods do not specify the shape of the probability density. The **histogram** is the simplest and best known non-parametric method for density estimation.

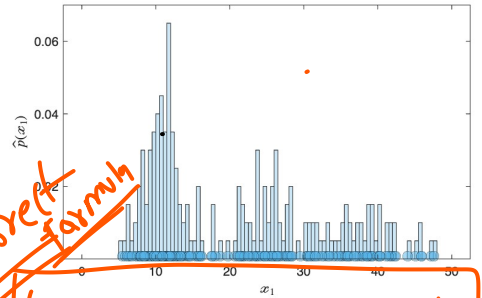
A histogram is built by dividing the feature space into equal-sized regions called **bins**. The density is approximated by the fraction of samples that fall within each bin.

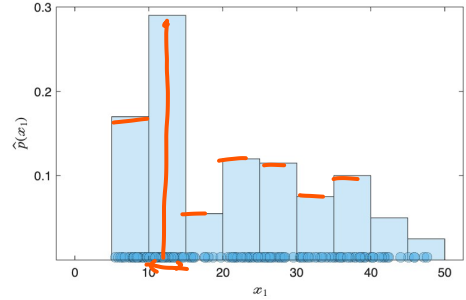




The histogram

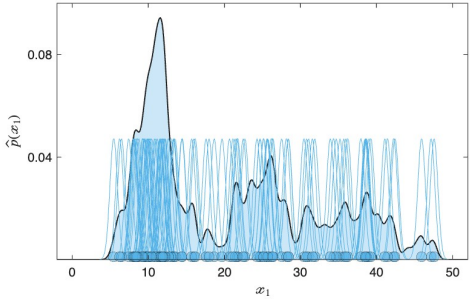
If bins are small, the estimated density will be **spiky**. If they are big, the estimated density will be **flat** and the underlying structure will be lost.

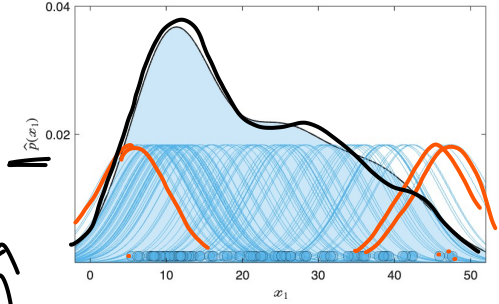




Kernel methods

Kernel methods proceed by building an **individual density** around each sample first and then **combining** all the densities together. Individual densities have the same shape (the **kernel**), for instance a Gaussian.





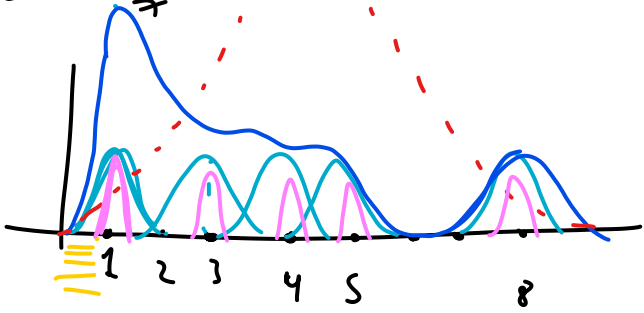
Kernel methods: Example

KDE

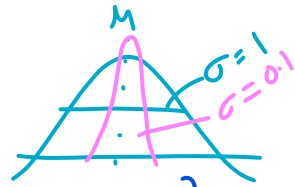
$x = [1, 3, 4, 5, 1, 1, 1, 8]$

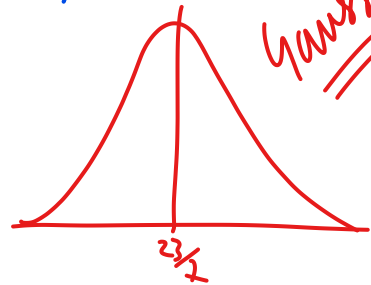
$\mu = 23/7$

$\sigma = \frac{1}{7} \Sigma (x-\mu)^2 =$ 2.



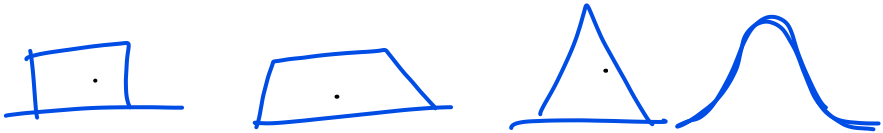
$e^{-\frac{(x-\mu)^2}{\sigma^2}}$



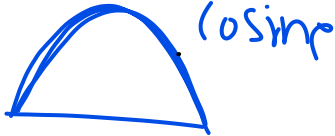


Gaussian

Summary



- Non-parametric methods for density estimation do not assume any specific shape for the probability density.
- They contain hyperparameters that need to be specified, such as the bin size or the type of kernel.
- The histogram provides a discrete probability density, whereas kernel methods produce a smooth one.
- Beware of the curse of dimensionality: histograms can end up having one sample per bin; individual distributions in kernel methods might end up being isolated from one another.



Agenda

Probability densities

Non-parametric methods

Parametric density estimation

Applications

Parametric density estimation

Parametric approaches **specify the shape** of the probability density. The problem of density estimation consists of **estimating its parameters**.

There are many available models, including:

- The Gaussian distribution, usually denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Log-normal distribution.
- Uniform distribution.
- Gamma distribution.

And many, many more.

The Gaussian distribution

The Gaussian or normal distribution $N(\mu, \sigma)$ is defined by two parameters μ and σ describing its **location** and **width**.

$x = [1.1, 3.4, 1.5, 6.7, 8]$

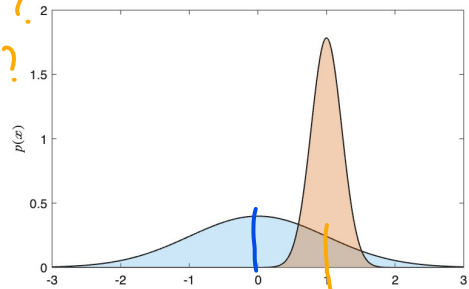
$\mu = ?$

$\sigma = ?$

Its mathematical expression in a 1D attribute spaces is:

$$p(x_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1-\mu}{\sigma}\right)^2}$$

μ is known as the **mean** and σ is the **standard deviation**. The value σ^2 is known as the **variance**.



$\mu = 0$, $\sigma = 5$

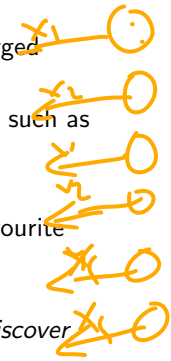
$\mu = 1$, $\sigma = 1$

The Central Limit theorem

The Central Limit Theorem (CLT) states that the **sum** of a large number of independent, random quantities has a **Gaussian** distribution.

This explains why, for instance:

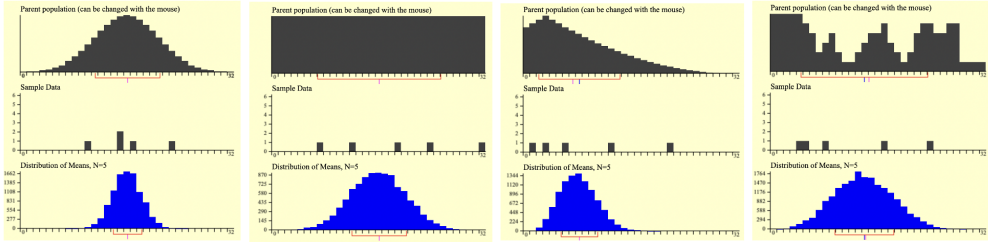
- Noise in electronic circuits is Gaussian, as it is the result of many individual charged particles subjected to random, thermal motion.
- Some physical traits are Gaussian, as they result from a large number of factors, such as genes, nutrition, etc.



The CLT is perhaps the **main reason** why the Gaussian distribution is one of our favourite density models.

(*There is another implication of the CLT: if your data is Gaussian, there is little to discover. Sounds interesting? Read about Information Theory*)

The Central Limit theorem



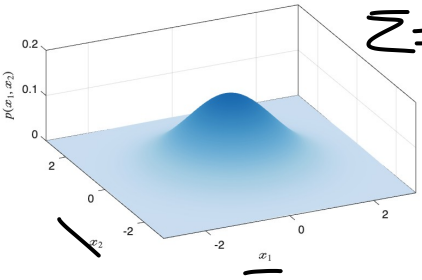
Demo Link: `https://onlinestatbook.com/stat_sim/sampling_dist/index.html`

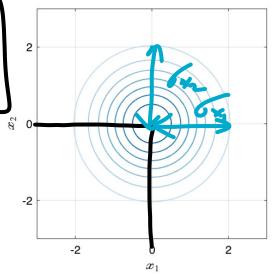
The multivariate Gaussian distribution

The Gaussian distribution can be extended to 2D, 3D... attribute spaces:

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

Where $\boldsymbol{x} = [x_1, \ldots, x_P]^T$ contains all the attributes. The mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ describe the **position** and **shape** of the distribution.

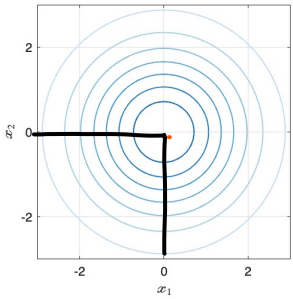




The multivariate Gaussian distribution: Example

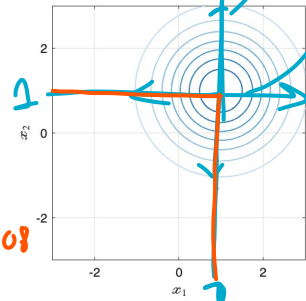
$\boldsymbol{\mu} = [0, 0]^T$

$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$



$\boldsymbol{\mu} = [1, 1]^T$

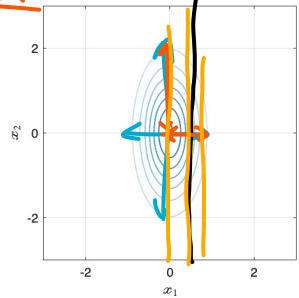
$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

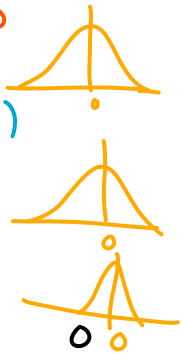


The multivariate Gaussian distribution: Example

$$\boldsymbol{\mu} = [0, 0]^T$$

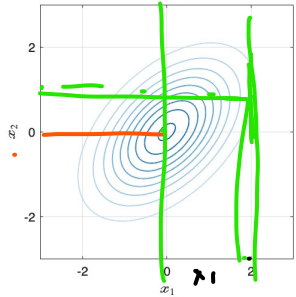
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & 0 \\ 0 & 2 \end{bmatrix}$$





$$\boldsymbol{\mu} = [1, 1]^T$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



Gaussian distributions and independent attributes

There is one more remarkable **property** of gaussian distributions. Given a Gaussian distribution $p(x_1, x_2)$, the following statements are equivalent:

- Attributes x_1 and x_2 are **independent** (e.g. we cannot predict the value of one based on the other).
- The covariance matrix Σ is **diagonal**.
- $p(x_1, x_2)$ can be obtained as the **product of the marginal densities** $p(x_1)$ and $p(x_2)$, which are themselves Gaussian.

This can be extended to higher dimensional attribute spaces. Remember this property when we talk about **Naive Bayes** approaches.

Gaussian distributions and independent attributes: Question

- Number of parameters
- Connection to **Naive Bayes** approaches

$N=10 \Rightarrow 20$

x_1 & x_2 indep. $N=2$

$P(x_1, x_2) = P(x_1)P(x_2)$

$P(x_1)$ | 2 parameters (μ, σ)

$P(x_2)$ | 2 parameters (μ, σ)

4 parameters

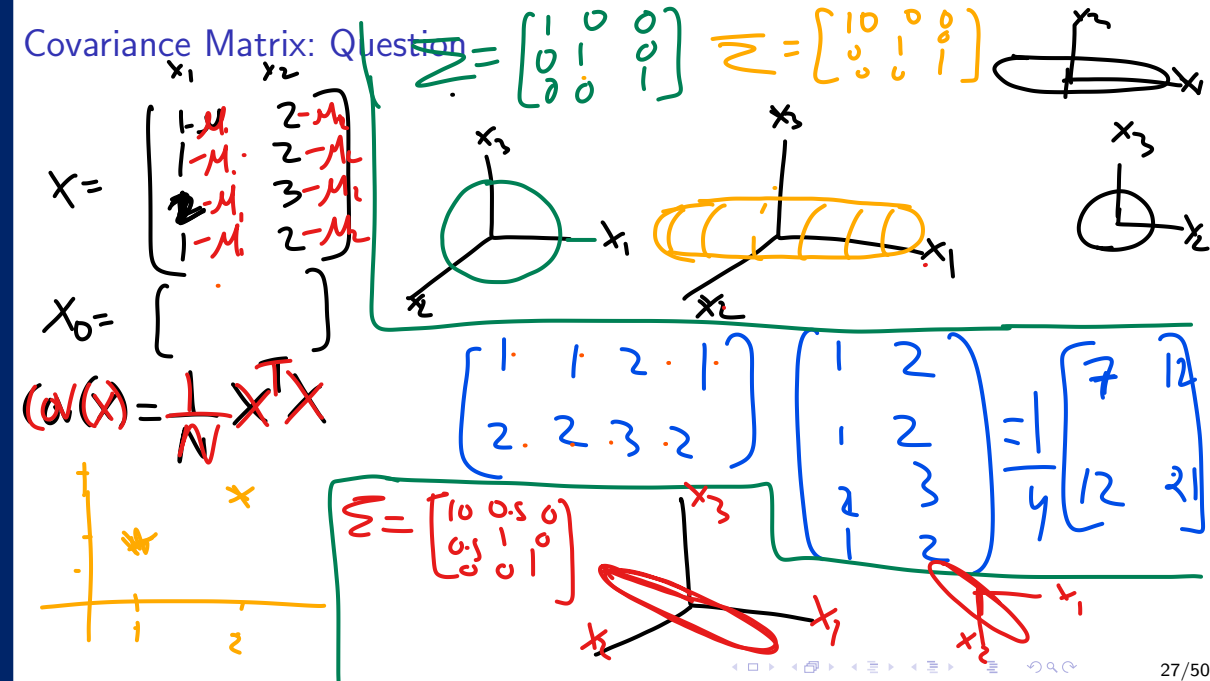
$N=2$ | $N=10$ 100+10 110

x_1 & x_2 $\neq$ indep

$P(x_1, x_2) =$ | $\mu = \begin{bmatrix} \cdot \\ \cdot \end{bmatrix}$ | $\Sigma = \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \end{bmatrix}$

6 parameters

Covariance Matrix: Question



Gaussian distribution: Estimation

If we are given a dataset consisting of N samples $\boldsymbol{x}_i$, the parameters of a Gaussian distribution can be estimated using **maximum likelihood** approaches.

In 1D attribute spaces, the parameters μ and σ^2 can be estimated as:

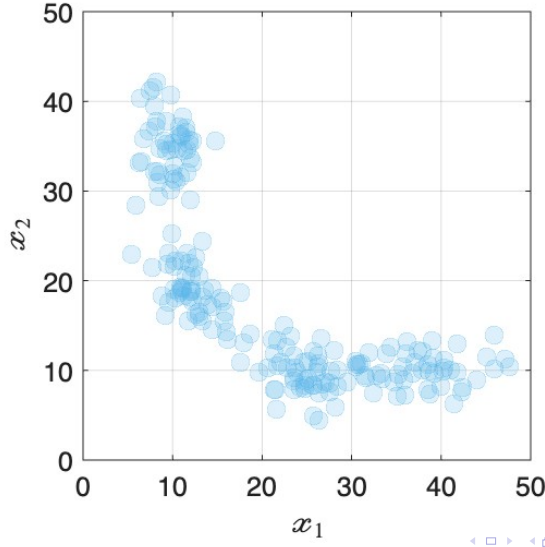
$$\hat{\mu} = \frac{1}{N}\sum_i x_i, \quad \text{and} \quad \hat{\sigma^2} = \frac{1}{N}\sum_i (x_i - \hat{\mu})^2$$

In higher dimensional spaces, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated as

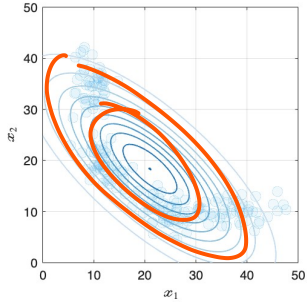
$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \boldsymbol{x}_i \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T$$

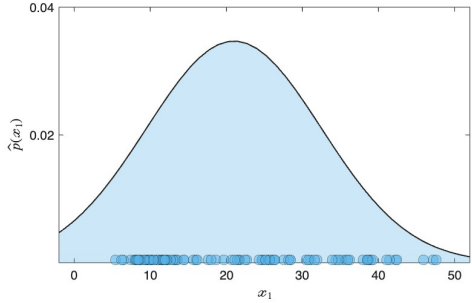
What is the total number of parameters that we are estimating?

Gaussian distribution: Estimation



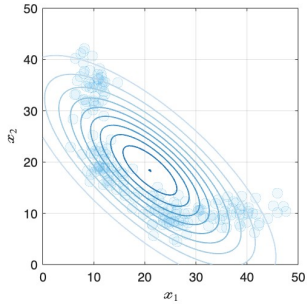
Gaussian distribution: Estimation



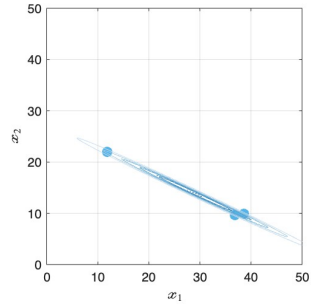


Gaussian distribution: High dimensionality

High dimensionality is always problematic and leads to overfitting. Constraints on the covariance matrix and regularisation techniques can be used to stabilise the solution.



Gaussian distribution obtained using all the samples.



Gaussian distributions obtained using subsets of samples.

Mixture models

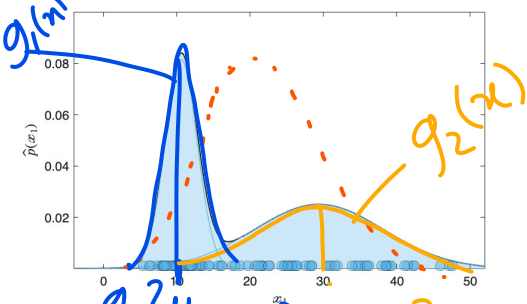
Datasets can exhibit **more than one mode** (*clumps*) for which single Gaussian densities are not suitable. In such cases, mixture densities such as **Gaussian Mixture Models** (GMM) constitute a convenient choice.

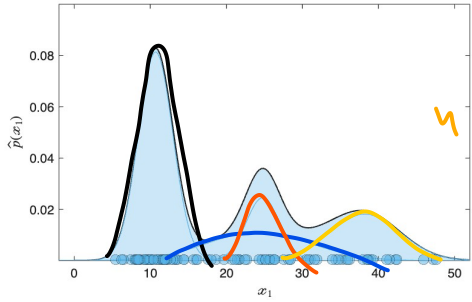
A GMM probability density is formulated as a **combination of Gaussian densities** $g_m(\boldsymbol{x})$ with their own mean $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$:

$$p(\boldsymbol{x}) = \sum_m g_m(\boldsymbol{x})\pi_m$$

where π_m are the mixing coefficients.

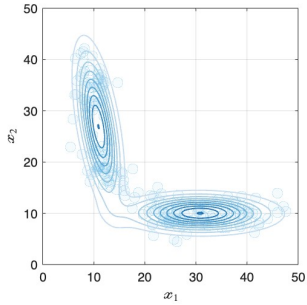
Mixture models: 1D Example



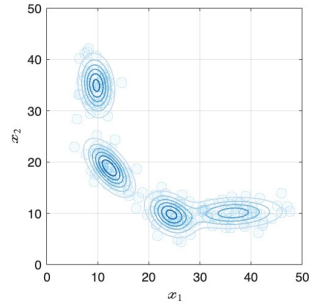


Mixture models: 2D Example

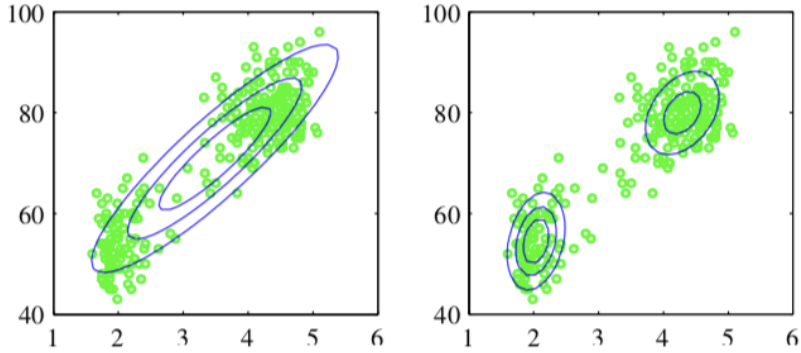
M=2



M=4

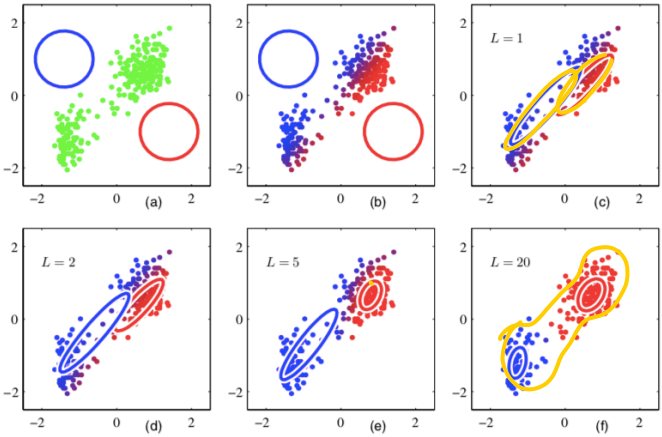


Mixture models: Example



Mixture models

The expectation-minimisation (EM) algorithm is an iterative process to fit a GMM to a dataset, similar to K-means.



Agenda

Probability densities

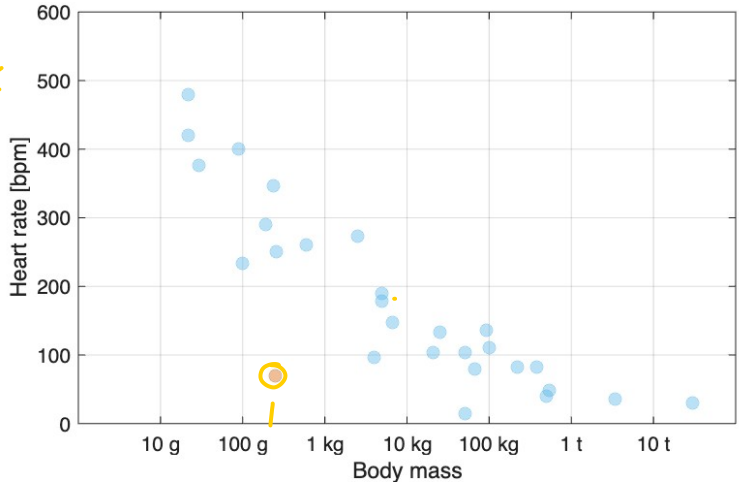
Non-parametric methods

Parametric density estimation

Applications

Back to the zoo

Can you spot a misbehaving animal?



Noise and outliers

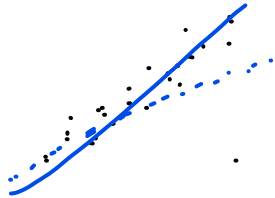
Anomaly Detection

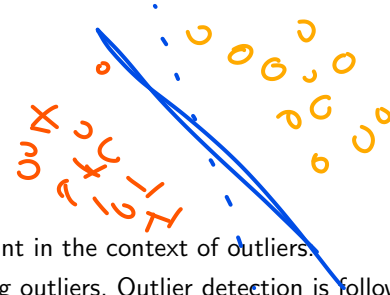
Samples extracted from the same population will always exhibit some level of **randomness** and deviate from the underlying **pattern**. Such deviations are known as **noise**.

Sometimes a sample can be so different that we doubt its deviation is just due to noise. We call these samples **outliers** or **anomalies**. Outliers are samples that belong to a different population altogether.

Detecting outliers is important for many **applications** (for instance, fraud detection or cybersecurity). Outliers can also have a **negative impact** if used during model training.

Outliers





There are two main families of methods that are relevant in the context of outliers:

- **Anomaly detection algorithms** aim at identifying outliers. Outlier detection is followed by actions that depend on the application, for instance, removal.
- **Robustness** is a design requirement that mitigates the impact of outliers. For instance, different cost functions can account differently for the cost associated to large deviations.

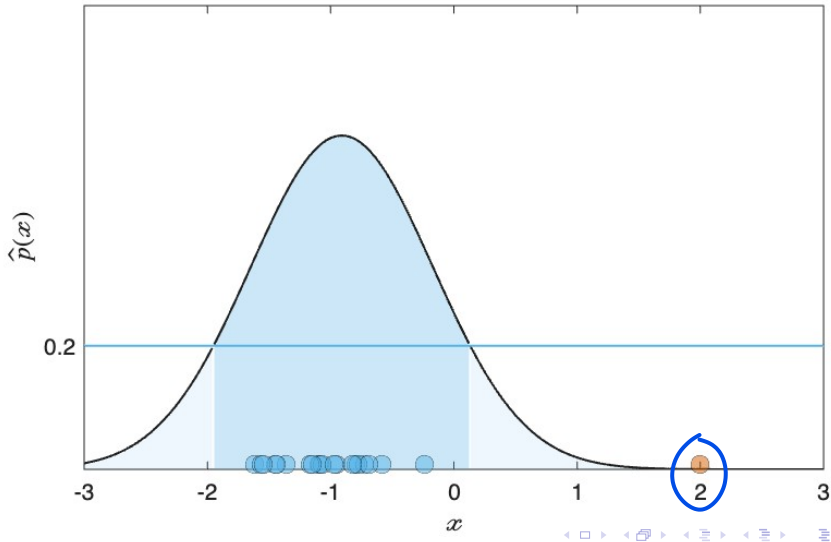
Basic anomaly detection algorithm

The main idea behind an anomaly detection algorithm is to quantify the probability of observing samples some distance away from the general pattern. If this probability is low, the sample is an anomaly.

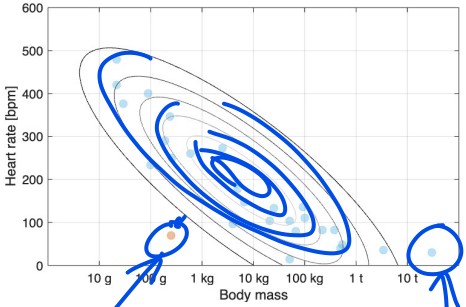
The main pattern is described by a probability density $p(\boldsymbol{x})$. If $p(\boldsymbol{x})$ is a multivariate Gaussian distribution, we proceed as follows:

- Estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the dataset.
- Agree on a threshold value T.
- If $p(\boldsymbol{x}_i) < T$, $\boldsymbol{x}_i$ is an anomaly.

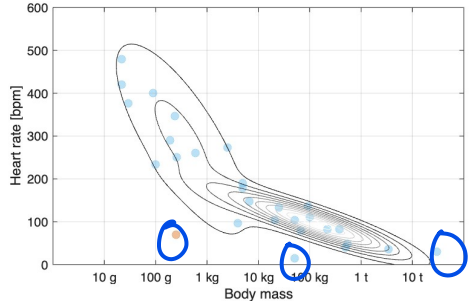
Basic anomaly detection algorithm: 1D Case



Basic anomaly detection algorithm



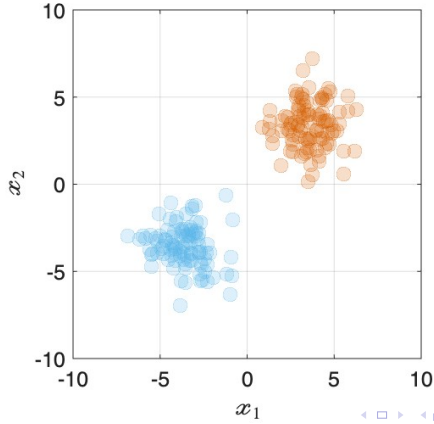
Single Gaussian.



Two Gaussians GMM.

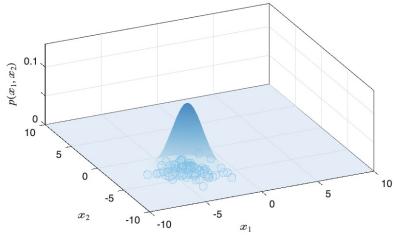
Classification: Estimating class densities

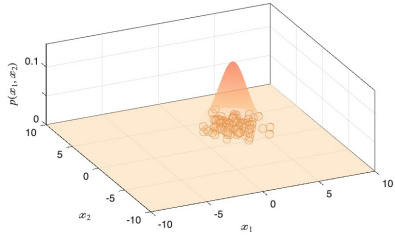
Classifiers that apply Bayes rule turn posterior probabilities into priors and class densities. A class density $p(\boldsymbol{x}|C)$ describes the distribution of samples in the predictor space for each class C.



Classification: Estimating class densities

Class densities are obtained using density estimation methods. We need to fit a probability distribution for each class **separately**.





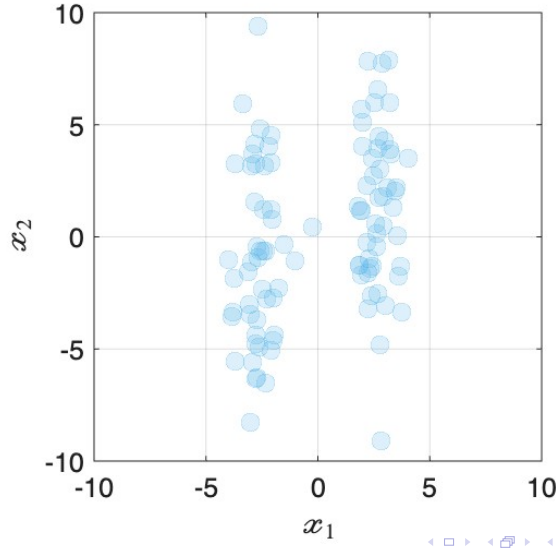
Naive Bayes classifiers

Gaussian distributions are the most popular choice for class densities. In high-dimensional scenarios, the total number of parameters is very large: for P predictors, we have P (mean) + P^2 (covariance) parameters.

Naive Bayes classifiers make the (naive) assumption that predictors are **independent**, hence a P-dimensional Gaussian distribution can be expressed as the product of its P **marginal distributions**.

As a result of this additional constraint, we need to obtain P (means) + P (variances) parameters, which reduces the risk of overfitting.

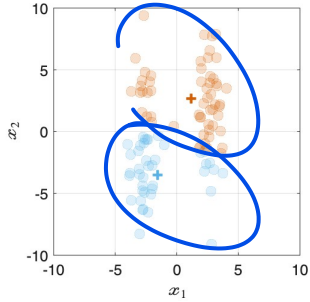
Clustering



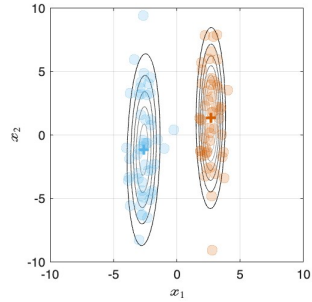
Clustering

K-means can be seen as a version of GMM fitting, where the Gaussian distributions have the same diagonal covariance matrix and hence clusters tend to be spherical.

GMM can be used as a clustering method that produces ellipsoidal clusters. First we fit K Gaussian densities and then we assign each sample to the most likely density.



K-Means



GMM

Summary

- **Probability densities** are models that allow us to calculate the probability of finding a sample in a region of the attribute space.
- **Non-parametric** methods do not assume any particular shape for the probability density, whereas **parametric** methods do.
- The Central Limit Theorem and other mathematical properties, make the **Gaussian distribution** one of the most popular choices.
- Probability densities can be used in many machine learning problems, such as **anomaly detection**, **classification** and **clustering**.

Queen Mary
University of London