School of Electronic Engineering and Computer Science
Queen Mary University of London

CBU5201 Principles of Machine Learning
Supervised learning: Classification II

Dr Chao Liu

Credit to Dr Jesús Requena Carrión

Oct 2023

The best diagnostic machine

As the hospital's lead data scientist, you are responsible for selecting the best diagnostic machine for a certain disease. You are presented three machines A, B and C, which **you test** using a group of patients whose diagnose you already know (test dataset).

The resulting accuracy of each machine after testing is shown below. Which one would you choose?

(a) Machine A: 6 % of correct diagnoses

(b) Machine B: 89 % of correct diagnoses

(c) Machine C: 56 % of correct diagnoses

Do I have the disease?

The prevalence of the disease is 0.1 %, i.e. we expect one individual out of 1000 to carry the disease.

Your diagnostic machine is such that:

- If you have the disease, the test is 100 % accurate.
- If you don't have the disease, the test is 95 % accurate.

You decide to take the medical test and the result is positive, the probability that you actually have the disease is:

(a) >97 %

(b) ≈ 50 %

(c) <3 %

Know your metrics!

The logistic model and kNN

We have introduced the **accuracy** and **error rate** as two convenient metrics to measure the quality of a classifier.

The following two approaches to build a classifier have been discussed:

- **Logistic model**: Trains a linear model using the likelihood or log-likelihood function as a quality metric during optimisation.
- **kNN**: Instance-based model that simply compares the proportion of samples of each class within a neighbourhood.

Note that neither of them seem to be defined based on the notions of accuracy and error rate. Do these notions play any role at all?

Agenda

The Bayes classifier

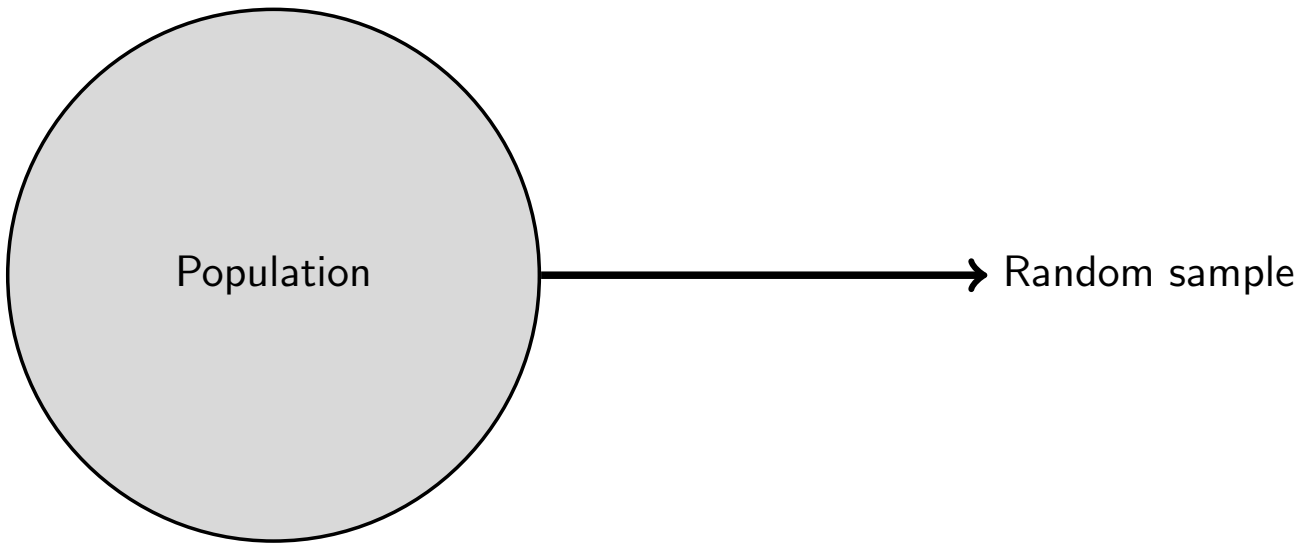
Using data to build posterior probabilities

Classification performance: Beyond accuracy

Summary

Which classifier has the highest accuracy?

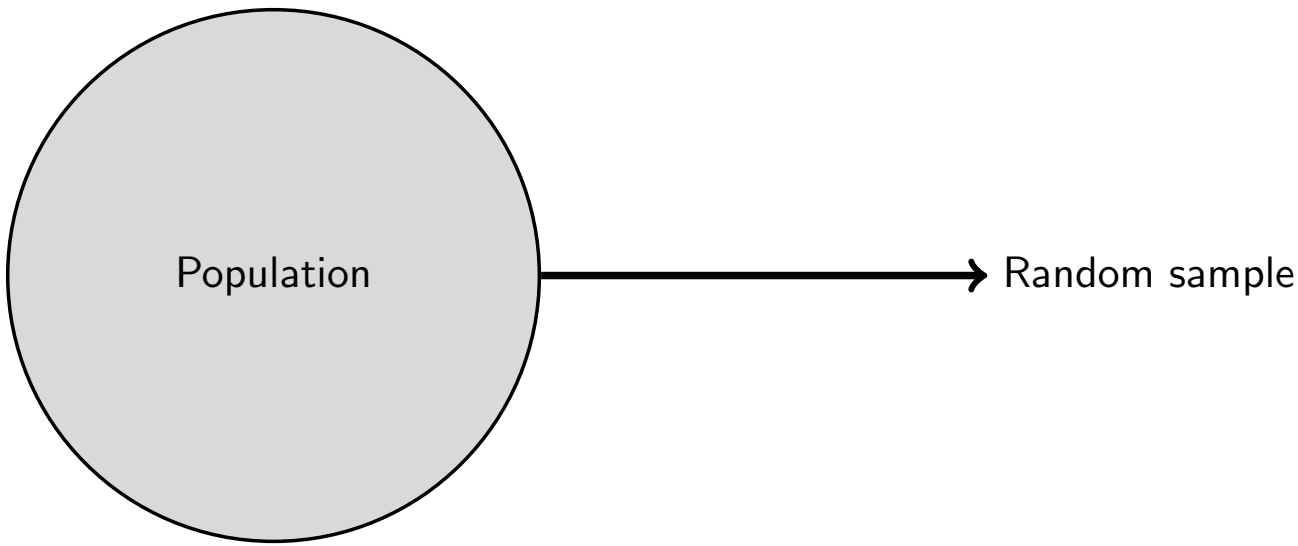
Consider a population consisting of individuals with an attribute y that can take on two values: o or o.



How do we determine whether a sample extracted randomly belongs to class o or o? How can we use any **available information** to achieve the **highest classification accuracy**?

The coin as a classifier

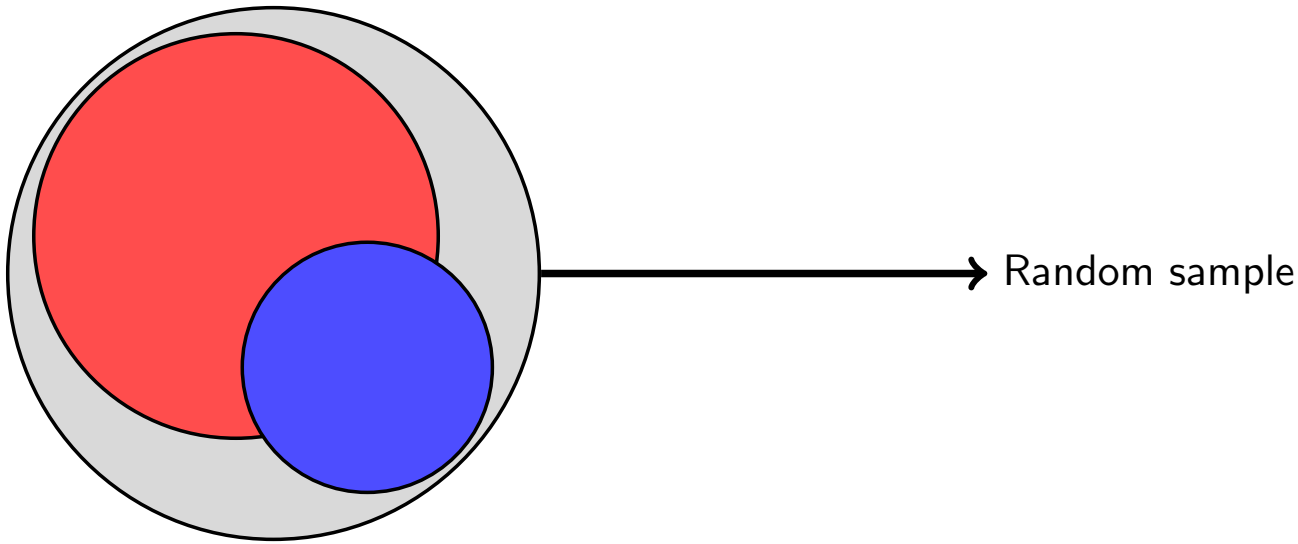
Without any additional information, we have no basis to decide which class the sample belongs to.



Flipping a o / o coin would be the best we can do (*Fifty-fifty*).

Prior probabilities

Now assume that we know that in 70 % of the population y = o, whereas in the remaining 30 % y = o.



In this scenario, we know the **class priors**, namely $P(y = \text{o}) = 0.7$ and $P(y = \text{o}) = 0.3$. We would **always** choose o.

Class densities

What if we have access to the value of **another attribute** x and know **how frequently** a value of x appears within each class?

For instance, assume we have the following insight:

- In $1/4$ of the o samples, $x = a$ and in $3/4$ $x = b$.
- In $2/3$ of the o samples $x = a$ and in $1/3$ $x = b$.

These are **class densities** and are expressed as follows:

$$p(x = a|y = \text{o}) = 1/4, \qquad p(x = b|y = \text{o}) = 3/4$$
$$p(x = a|y = \text{o}) = 2/3, \qquad p(x = b|y = \text{o}) = 1/3$$

Given a sample where $x = a$, we could compare $p(x = a|y = \text{o})$ and $p(x = a|y = \text{o})$ and decide that the sample belongs to o.

Posterior probabilities and the Bayes classifier

It turns out that the classifier that achieves the **highest accuracy** is the one that compares the **posterior probabilities**, defined as the probability that a sample belongs to a class given the value of its predictors, e.g.

$$P(y = \text{o}|x = a) \lessgtr P(y = \text{o}|x = a)$$

Do not confuse a posterior probability with a class density!

The classifier that uses the **true posterior probabilities** is called the **Bayes classifier**. The Bayes classifier uses the **odds ratio**:

$$\frac{P(y = \text{o}\ |\boldsymbol{x})}{P(y = \text{o}|\boldsymbol{x})} \lessgtr 1$$

If the ratio is greater than 1, the sample is o, if it is less, it is o. This is equivalent to assigning the sample to the most probable class.

Agenda

Machine learning as statistical learning

To achieve the highest classification accuracy, we need to know the posterior probabilities. The question is, *where do we get them from*?

Machine learning classifiers use datasets to **estimate the posterior probabilities**. For instance:

- **Logistic models** use the logistic function to build a posterior probability (the classifier's *certainty*).
- In **kNN** the posterior probabilities are estimated as the proportion of neighbours that belong to each class .

The estimated posterior probabilities will in general be different from the **true posterior probabilities** and hence these classifiers will never beat the Bayes classifier.

Bayes rule

If we happen to know the **priors** and the **class densities**, we can apply Bayes rule to obtain the **posterior probabilities** exactly:

$$P(y = \text{o} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = \text{o})P(y = \text{o})}{p(\boldsymbol{x})}, \quad P(y = \text{o} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = \text{o})P(y = \text{o})}{p(\boldsymbol{x})}$$

The odds ratio can be expressed using the priors and the class densities:

$$\frac{P(y = \text{o} \mid \boldsymbol{x})}{P(y = \text{o}|\boldsymbol{x})} = \frac{p(\boldsymbol{x}|y = \text{o})P(y = \text{o})}{p(\boldsymbol{x}|y = \text{o})P(y = \text{o})} \lessgtr 1$$

Hence, the problem of building posterior probabilities is equivalent to the problem of building priors and class densities.

Bayes rule in machine learning

In machine learning, we can use **data** to estimate the **priors** and the **class densities**.

Using a dataset to estimate the **priors** is very easy, we simply need to count the number of samples belonging to each class:

$$P(y = \text{o}) = \frac{\#\ \text{o samples}}{\#\ \text{samples}}, \quad P(y = \text{o}) = \frac{\#\ \text{o samples}}{\#\ \text{samples}}$$

Building the **class densities** $p(\boldsymbol{x}|y = \text{o})$ and $p(\boldsymbol{x}|y = \text{o})$ is in fact an **unsupervised problem**.

For now, we will focus on the most common density, the **Gaussian density**.

Discriminant analysis

In **discriminant analysis**, we assume that the class densities are **Gaussian**. If there is one predictor x, the o class density is:

$$p(x|y = \text{o}) = \frac{1}{\sigma_\text{o}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_\text{o}}{\sigma_\text{o}})^2}$$

where μ_o is the **mean** and σ^2_o the **variance** of the Gaussian density.

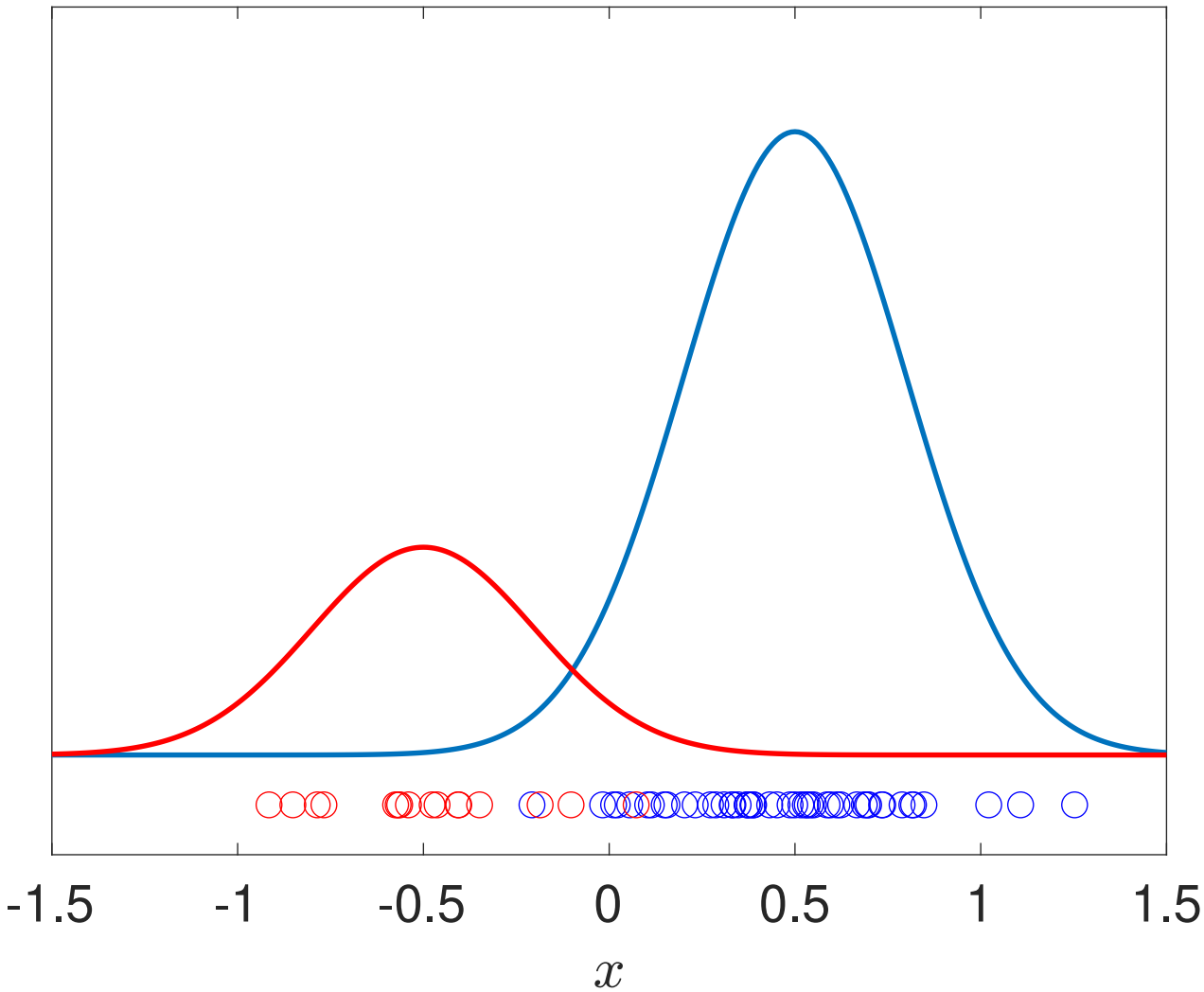
If there are K predictors, a Gaussian class density is expressed as:

$$p(\boldsymbol{x}|y = \text{o}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_\text{o}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_\text{o})^T \boldsymbol{\Sigma}_\text{o}(\boldsymbol{x}-\boldsymbol{\mu}_\text{o})}$$

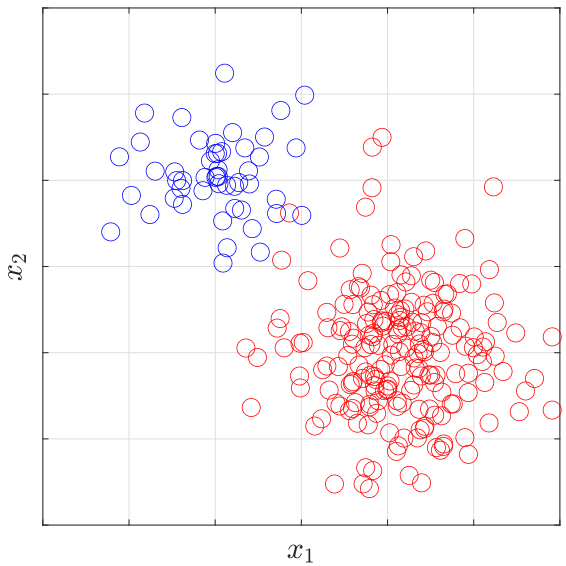
we $\boldsymbol{x} = [x_1, \ldots, x_K]^T$ contains all the predictors (note we have **not** prepended a 1), $\boldsymbol{\mu}_\text{o}$ is the **mean** and $\boldsymbol{\Sigma}_\text{o}$ is the **covariance matrix**.

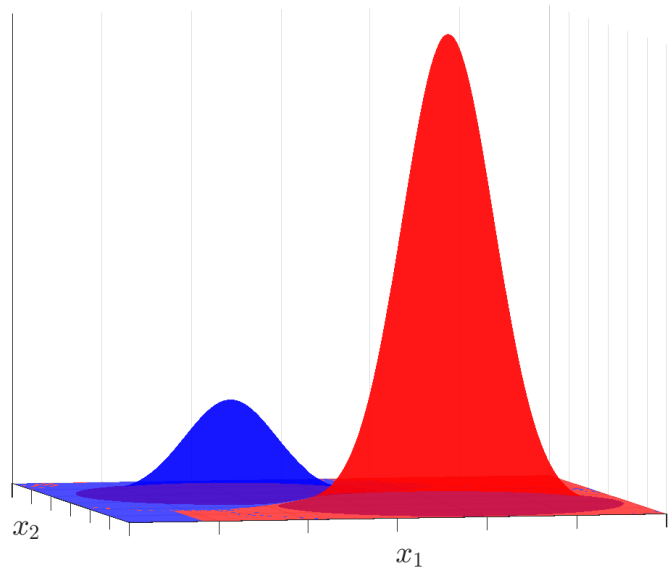
Similar expressions can be obtained for the o class densities.

Discriminant analysis: one predictor



Discriminant analysis: two predictors

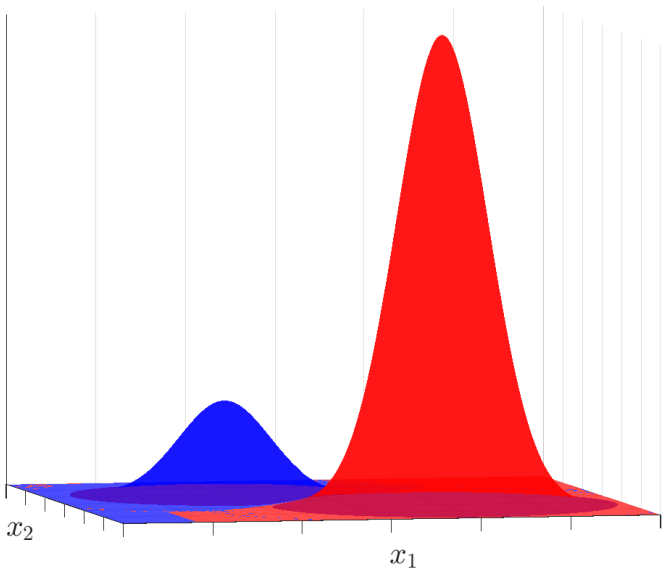


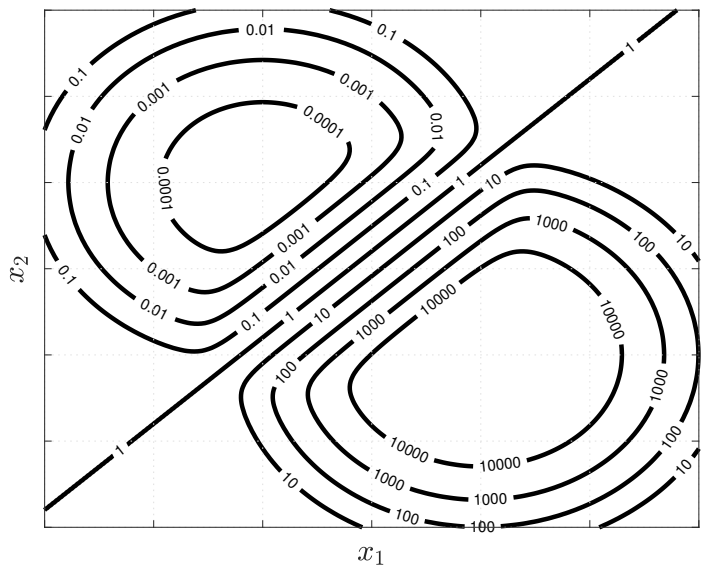


Linear and quadratic discriminant analysis

The boundary in discriminant analysis depends on the covariance matrices $\mathbf{\Sigma}_o$ and $\mathbf{\Sigma}_o$:

- If $\mathbf{\Sigma}_o = \mathbf{\Sigma}_o$ the boundary is linear. We call this scenario **linear discriminant analysis** (LDA).
- Otherwise, the boundary is quadratic. This is **quadratic discriminant analysis** (QDA).





Comparison of classifiers seen so far

- **Shape of boundaries**: Logistic regression and LDA build linear boundaries, QDA quadratic boundaries. kNN does not impose any particular shape.
- **Stability**: For a small number of samples, logistic regression can be very unstable, whereas DA approaches produce stable solutions.
- **Outliers**: Logistic regression is robust against samples which lie very far from the boundary, LDA and QDA can be affected.
- **Multiclass**: Multiclass problems can be implemented easily in dicriminant analysis.
- **Prior knowledge**: Can be easily incorporated following Bayesian approaches.