QUEEN'S UNIVERSITY BELFAST

SCHOOL OF
ELECTRONICS,
ELECTRICAL
ENGNIEERING AND
COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 – 2019/2020

Feature and sample in a dataset - prediction

Consider the following example

It is important for the bank to be able to **predict** in advance **the risk associated with a loan**, which is the probability that the customer will default and not pay the whole amount back.

In **credit scoring**, the bank calculates the risk given the amount of credit and **the information about the customer**. The **information about the customer** includes data we have access to and is relevant in calculating his or her financial capacity - namely, **income**, **savings**, **collaterals**, **profession**, **age**, **past financial history**, and so forth. The bank has a record of past loans containing such customer data and whether the loan was paid back or not.

From this data of particular applications, the aim is **to infer a general rule** coding **the association between a customer's attributes and his risk**.

A machine learning system fits a model to the past data to be able to calculate the risk for a new application and then decides to accept or refuse it accordingly.

Originated from "Introduction to Machine Learning", Ethem Alpaydin, MIT Press, 2004

Prediction problem

- It is important for the bank to be able to **predict** in advance **the risk associated with a loan**, which is the probability that the customer will default and not pay the whole amount back.

- From a data of particular applications, the aim is to infer a general rule coding **association** between a **customer's attribute (features)** and **his/her risk**.

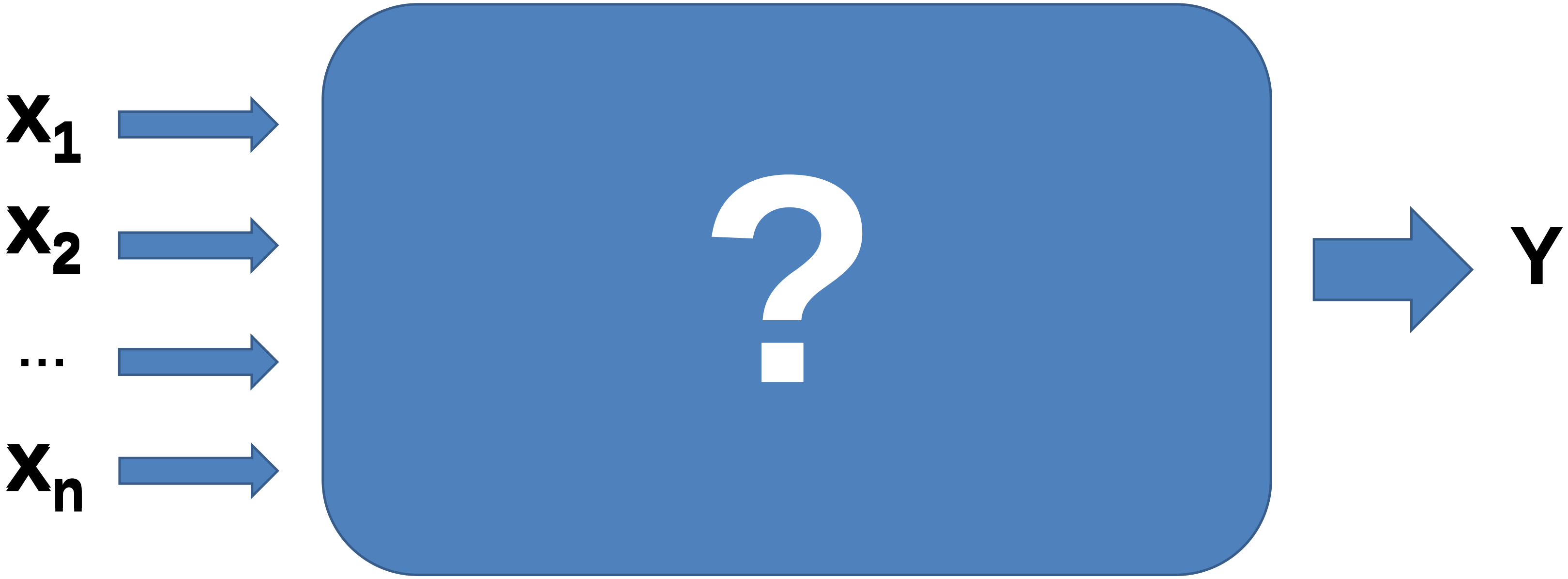
- This is an example of a ***classification*** problem where there are two classes: low-risk and high-risk customers. The information about a customer makes up the ***input*** to the classifier whose task is to assign the input to one of the two classes.

Prediction problem – *discriminant*

X_1: Income

X_2: Savings

Y: Low-risk or high-risk



- This is an example of *discriminant*; it is a function that separates the examples of different classes.
- **Discriminant analysis**

What is a sample and a feature in this dataset?

#	Income	Savings	Collaterals	Profession	age	…
1	£25K	2K	Yes	…	24	…
2	£35K	5K	No	…	30	…
3	£5K	0.5K	Yes	…	26	…
…	…	…	…	…	…	..
17	£45K	10K	No	…	40	…

Features

17 samples

?

What is a sample and a feature in this dataset?

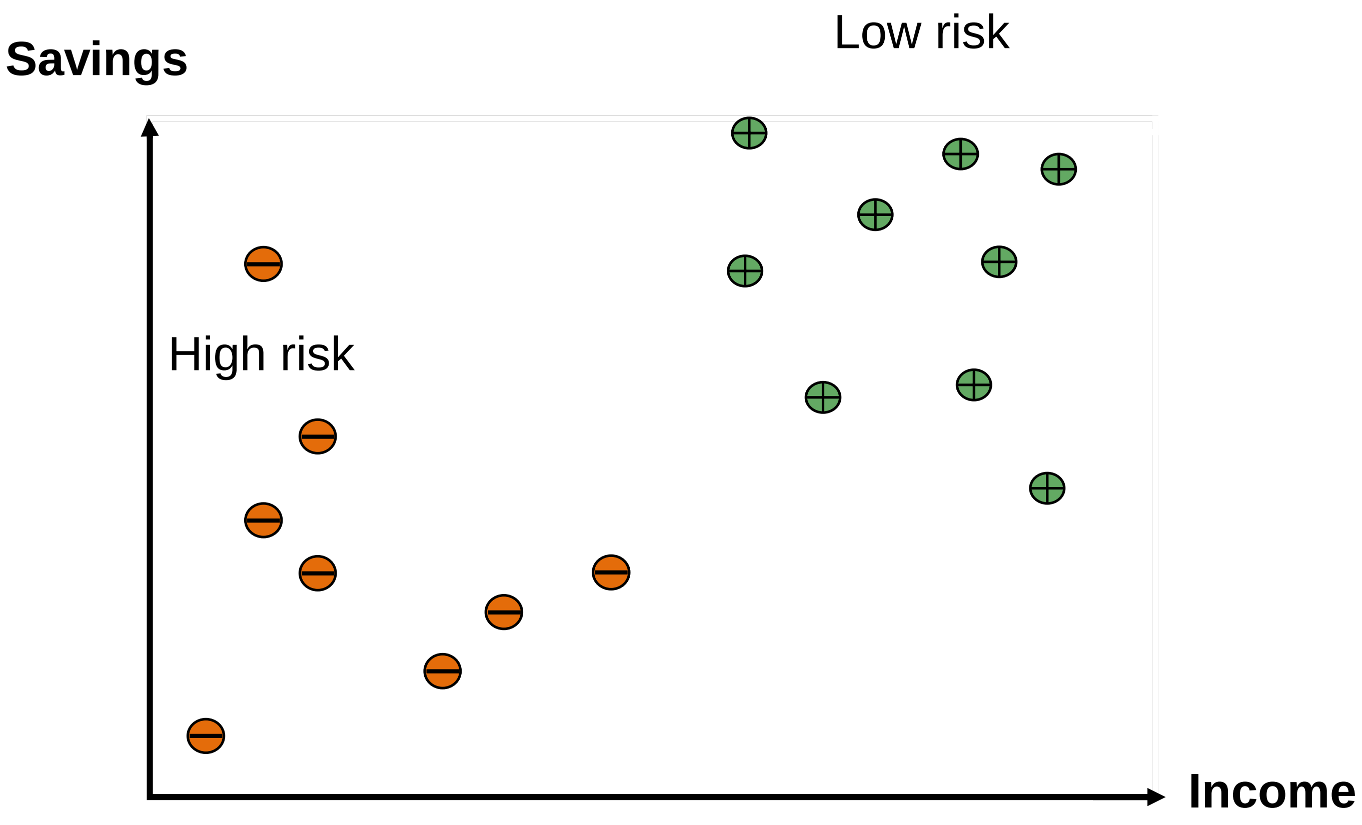
#	Income	Savings	Collaterals	Profession	age	…
1	£25K	2K	Yes	…	24	…
2	£35K	5K	No	…	30	…
3	£5K	0.5K	Yes	…	26	…
…	…	…	…	…	…	..
17	£45K	10K	No	…	40	…

Features

?

17 samples

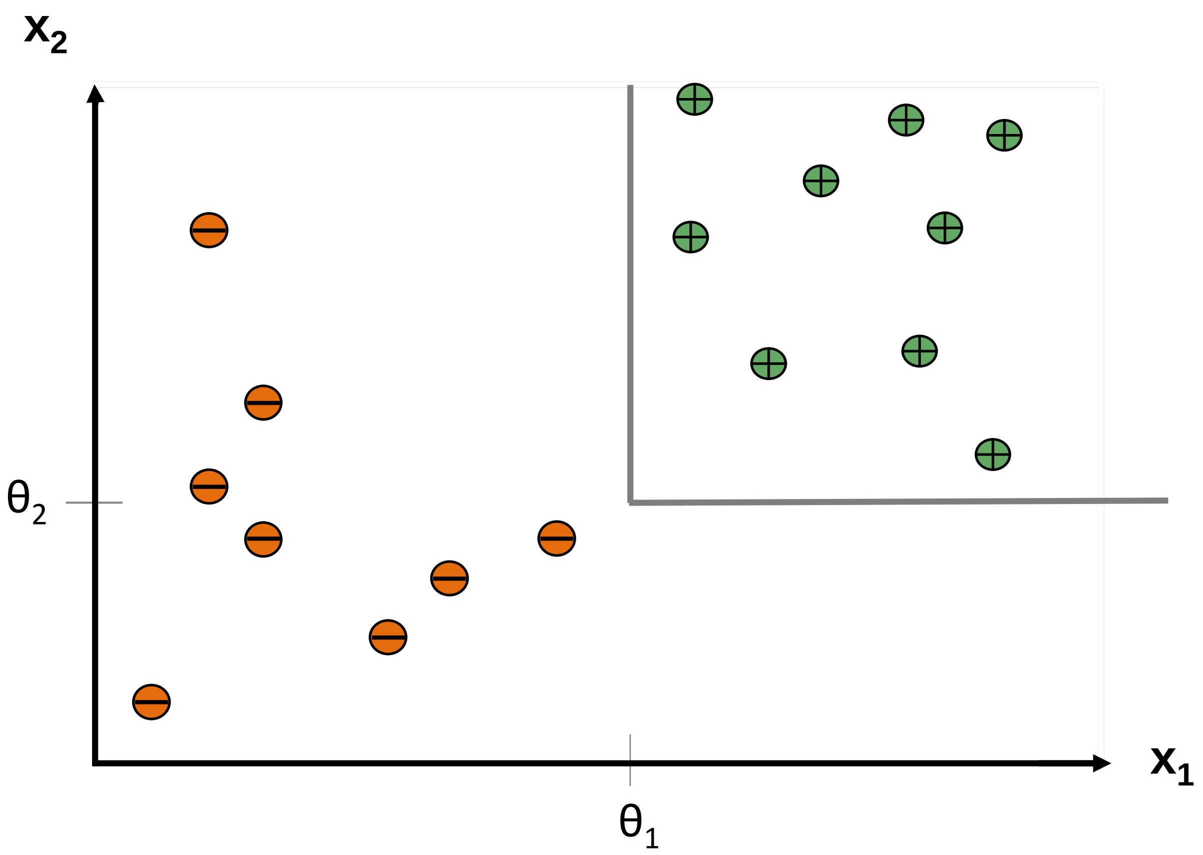
What is a sample and a feature in a dataset?



#	Income	Savings	Collaterals	Profession	age	…
1	£25K	2K	Yes	…	24	…
2	£35K	5K	No	…	30	…
3	£5K	0.5K	Yes	…	26	…
…	…	…	…	…	…	..
17	£45K	10K	No	…	40	…

This figure illustrates an example of a dataset. Each circle corresponds to one **data instance** with input values in the corresponding axes. **For simplicity**, only two customer **attributes or features**, income and savings, are taken **as input** and the two classes are low-risk (‘+’) and high-risk (‘−’).

Prediction problem



IF $x_1 > \theta_1$ and $x_2 > \theta_2$ THEN low-risk ELSE high-risk

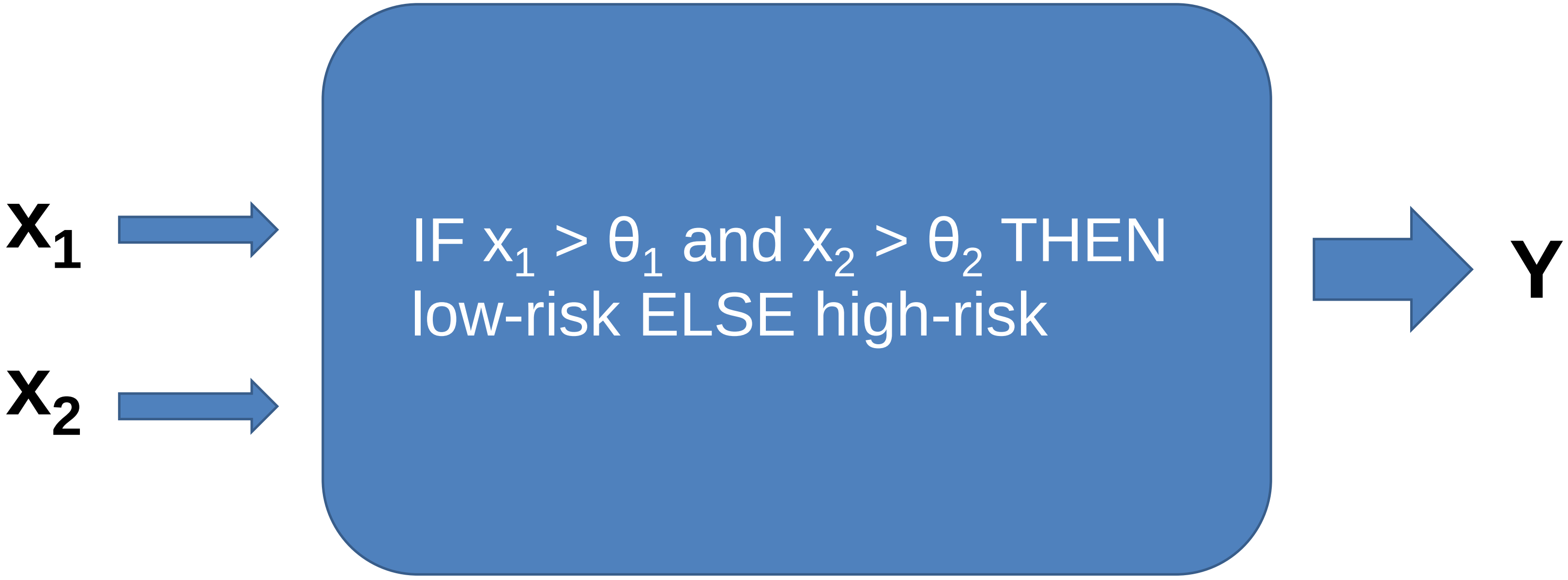
Once we have a rule like this that fits the past data, if the future is similar to past, then we can make correct predications for new instances.

Prediction problem – *discriminant*

X_1: Income

X_2: Savings

Y: Low-risk or high-risk



x_1

x_2

IF $x_1 > \theta_1$ and $x_2 > \theta_2$ THEN low-risk ELSE high-risk

Y

- This is an example of *discriminant*; it is a function that separates the examples of different classes.
- Discriminant analysis

Question & Discussion 1.1 (QD.1.1)

IF $x_1 > \theta_1$ and $x_2 > \theta_2$ THEN low-risk ELSE high-risk

- Think about all the input parameters which may affect the accuracy of the result from this function (i.e., model)?

- We will discuss about this question in our **Discussion Forum (Page)**. Please be involved and participate in our discussion!