

# Introduction to Machine Learning

Prof. Alessandro Lucantonio

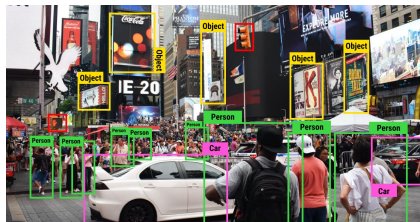
Aarhus University

24/10/2023

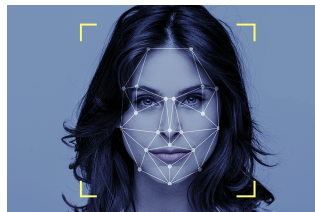
# What is Machine Learning?

- ▶ Arthur Samuel (1959). Machine learning is a “Field of study that gives computers the ability to learn without being explicitly programmed”.
- ▶ Tom Mitchell (1998). “A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ ”.

# Some applications - Image recognition



(a)



(b)

Two examples of image recognition:

(a) Labelling different entities in a given image.

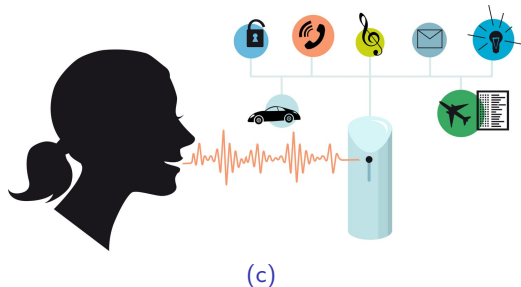
(b) Face recognition.

## Some applications - Self-driving cars



## Some applications - Speech and voice recognition

*Speech recognition* (or transcription) involves recording a speech using a recording device. The audio is then converted into a set of words stored digitally within the device or program. Instead, the purpose of *voice recognition* is to identify the person who is speaking.



(c)



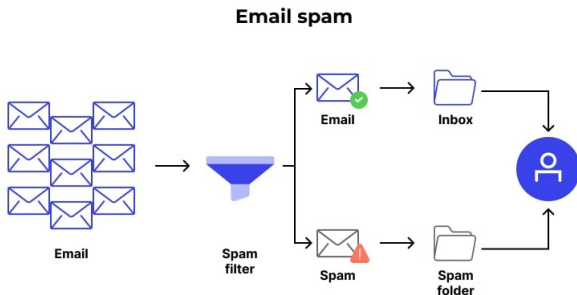
(d)

(c) A general idea of speech recognition.

(d) Google's voice assistant provide individualized responses (e.g. calendar updates or reminders) only to the users who trained the assistant to recognize their voices.

# Some applications - Email spam filtering

Task: determine if a given email is spam or not.



## Some applications - Learning how to play games



“AlphaGo is the first computer program to defeat a professional human Go player, the first to defeat a Go world champion, and is arguably the strongest Go player in history.”

More info: <https://www.deepmind.com/research/highlighted-research/alphago>

# Data

A **sample** (or example) is a collection of **attributes** (or features) that have been quantitatively measured from some object or event that we want the ML system to process.

A **dataset** is a collection of many samples.

Case	Attributes				Decision
	Length	Height	Width	Weight	Quality
1	4.7	1.8	1.7	1.7	high
2	4.5	1.4	1.8	0.9	high
3	4.7	1.8	1.9	1.3	high
4	4.5	1.8	1.7	1.3	medium
5	4.3	1.6	1.9	1.7	medium
6	4.3	1.4	1.7	0.9	low
7	4.5	1.6	1.9	0.9	very-low
8	4.5	1.4	1.8	1.3	very-low

We will deal with two types of attributes:

- ▶ Continuous = real numbers.
- ▶ Categorical (or discrete) = integers.



# Data - Attributes

Date	Location	MinTemp	MaxTemp	Rainfall	Humidity9am	Humidity3pm	Pressure9am	Pressure3pm	Temp9am	Temp3pm	RainToday	RainTomorrow
01/12/2008	Albury	13.4	22.9	0.6	71	22	1007.7	1007.1	16.9	21.8	No	No
02/12/2008	Albury	7.4	25.1	0	44	25	1010.6	1007.8	17.2	24.3	No	No
03/12/2008	Albury	12.9	25.7	0	38	30	1007.6	1008.7	21	23.2	No	No
04/12/2008	Albury	9.2	28	0	45	16	1017.6	1012.8	18.1	26.5	No	No
05/12/2008	Albury	17.5	32.3	1	82	33	1010.8	1006	17.8	29.7	No	No
06/12/2008	Albury	14.6	29.7	0.2	55	23	1009.2	1005.4	20.6	28.9	No	No
07/12/2008	Albury	14.3	25	0	49	19	1009.6	1008.2	18.1	24.6	No	No
08/12/2008	Albury	7.7	26.7	0	48	19	1013.4	1010.1	16.3	25.5	No	No
09/12/2008	Albury	9.7	31.9	0	42	9	1008.9	1003.6	18.3	30.2	No	Yes
10/12/2008	Albury	13.1	30.1	1.4	58	27	1007	1005.7	20.1	28.2	Yes	No
11/12/2008	Albury	13.4	30.4	0	48	22	1011.8	1008.7	20.4	28.8	No	Yes
12/12/2008	Albury	15.9	21.7	2.2	89	91	1010.5	1004.2	15.9	17	Yes	Yes
13/12/2008	Albury	15.9	18.6	15.6	76	93	994.3	993	17.4	15.8	Yes	Yes

. . . . .  
. . . . .  
. . . . .

**Figure:** “MinTemp” is an example of continuous attribute, while “RainToday” is an example of categorical attribute (Yes = 1, No = 0).

# Data - Preprocessing

Preprocessing: preparing and manipulating the dataset to make it suitable for the model.

- ▶ **Data cleaning.** Identifying and correcting inconsistencies in the data.
- ▶ **Data transformation.** Converting the data into a suitable format
- ▶ **Data reduction.** Reducing the size of the dataset preserving as much as you can the important information
- ▶ **Data normalization.** Scaling the data to a common range (usually between 0 and 1 or -1 and 1).

# Tasks - Supervised learning

The **task** is the purpose of the application.

**Supervised learning** consists of tasks where each sample is associated with a **target** (also called label).

- ▶ **Regression.** Predict a continuous dependent variable.  
Example: Predict the price of a house given its size.
- ▶ **Classification.** Predict which categories/classes the input belongs to (categorical data).  
Example: Given an e-mail, predict if it is spam or not (*binary classification*).

# Tasks- Unsupervised learning

**Unsupervised learning** tasks experience a dataset containing only features, but not a supervision signal (**unlabeled data**).

Examples:

- ▶ Take a collection of essays and find a way to automatically group them based on word frequency, sequence length, page counts etc.
- ▶ Recommender systems. Automatically provide suggestions for an item that is most pertinent to a particular user.

# Unsupervised learning - an example

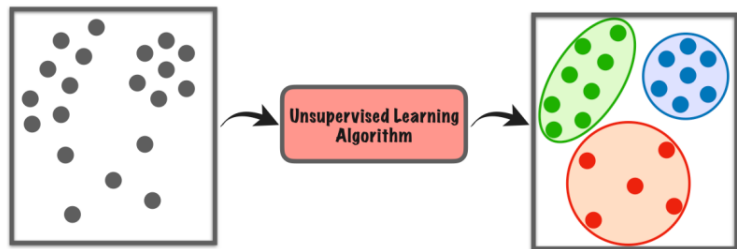


Figure: A clustering example.

# Model

A **model** is an abstract representation of a system using mathematical concepts. In the present context, the model operates on data to solve the task.

**Hypothesis:** a candidate model for the task.

**Hypothesis space:** the class of models that the learning algorithm (see later) can produce.

**No free lunch theorem:** averaged over *all* possible data-generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points. However, we can design algorithms that perform well on specific kinds of probability distributions that we encounter in real-world applications.

# Learning algorithm

A machine learning algorithm, in brief **learning algorithm**, is an algorithm that is able to learn from data.

To evaluate the abilities of a learning algorithm, we must design a quantitative measure of its performance. Usually this performance measure  $P$  is specific to the task being carried out by the system.

The performance measure is called **cost function** (or *loss function*).

Example: a performance measure for the classification tasks is the **accuracy**. Accuracy is just the proportion of examples for which the model produces the correct output.