Introduction to Machine Learning

Prof. Alessandro Lucantonio

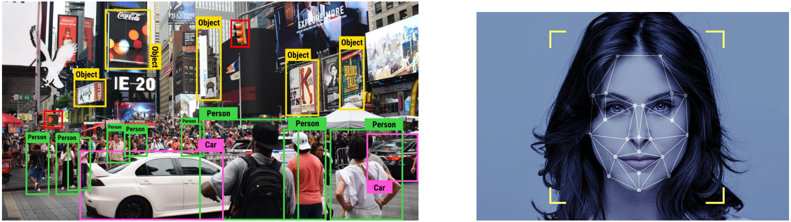
Aarhus University - Department of Mechanical and Production Engineering

?/?/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”.

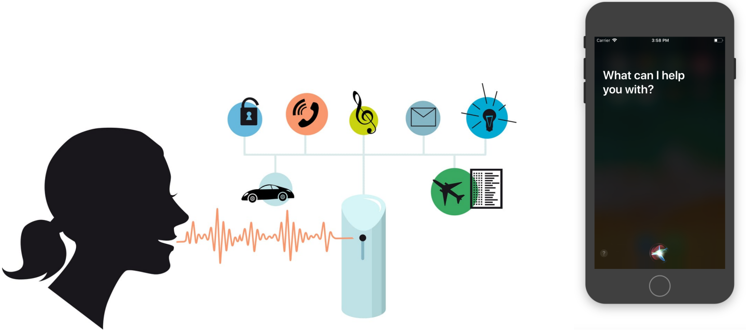
Some applications - Image recognition



(a) (b)

Two examples of image recognition.
(a) Labelling different entities in a given image.
(b) Face recognition (as in our smartphones).

Some applications - Speech and voice recognition



Carrier 3:58 PM

What can I help you with?

(c)

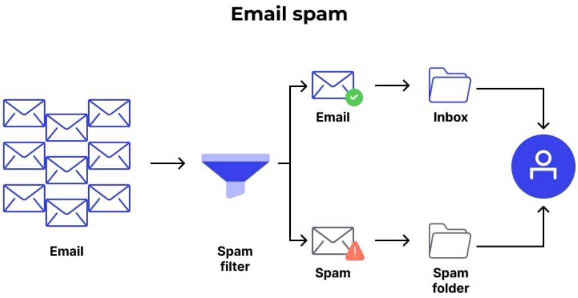
(d)

Two examples of speech and voice recognition.
(c) A general idea of speech recognition.
(d) Apple Siri.

Some applications - Self driving cars

Using e.g. image recognition, companies are building self-driving cars increasingly efficient.

Some applications - Email spam filtering

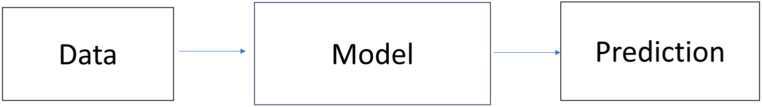


Determine if a given email is spam or not.

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

ML road map



Data

Model

Prediction

-How much data?
- Should I *pre-process* my data?
-Are my data *balanced*?

-What is the *task*?
-Which *learning algorithm* will I use?
-How to *validate my model?*

What is the performance on
unseen data?

Data

Data captures the structure of the problem.

The data instances are called **samples** and each sample has different **attribute** values. The kind and the number of attribute is common for each sample. We will see two types of attributes

- Continuous attributes (real numbers).
- Categorical (or discrete) attributes (integers).

Preprocessing: prepare and manipulate data to make them suitable for the model.

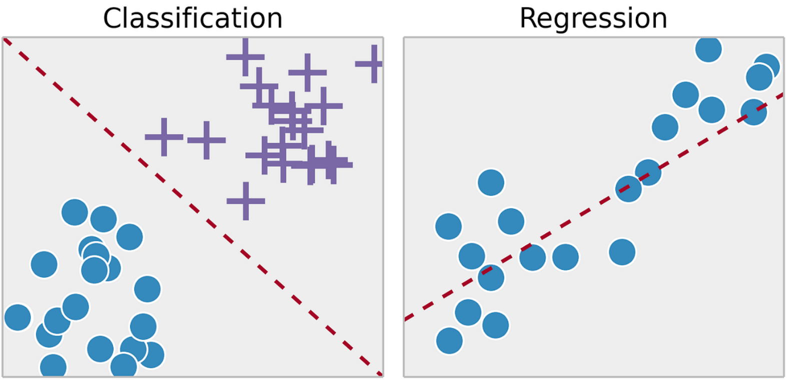
Tasks - Supervised Learning

The task is the purpose of the application.

In **supervised learning**, a dataset of input-output relations is provided. The learning is supervised because we already know how the current looks like.
Two type of supervised learning problems:

- **Regression**. Predict results within a continuous output.
  Example: Predict the price of an house given its size.
- **Classification**. Predict results within a discrete output (categorical data).
  Example: Given an email, predict if it is spam or not (*binary classification*)

Supervised Learning - an example



Tasks- Unsupervised Learning

In **unsupervised learning**, we have no idea how the output looks like (unlabeled data). We have to derive structure and different relationships from 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

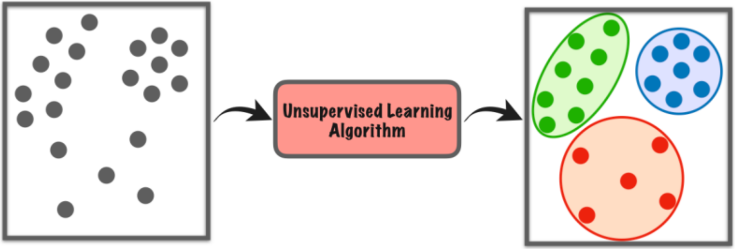


Figure: A clustering example.

Model

Main goal: describe data relationships through a language

Hypotheses: a candidate model for the task.
Hypotheses space: the class of hypotheses that the learning algorithm (see later) can produce.

No free lunch theorem (idea): it does not exist a universal best learning method.

Learning algorithm

Based on: data, task and model.

Search method: find the best hypothesis in the hypothesis space.

Best hypothesis means "minimum error". The error function is called *cost function* (or *loss function*).

Crucial: the aim is *generalize* and NOT *fitting*. We want to avoid models which performs extremely good on our data and poorly on unseen data (overfitting).