

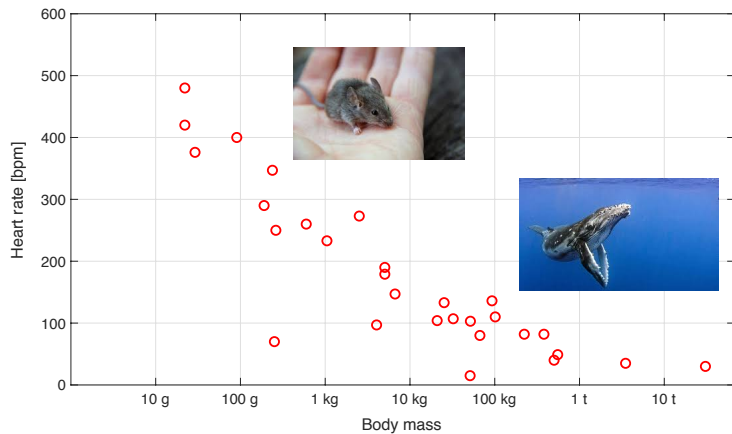
SCHOOL OF ELECTRONIC ENGINEERING AND COMPUTER SCIENCE
QUEEN MARY UNIVERSITY OF LONDON

ECS7020P Principles of Machine Learning Introduction

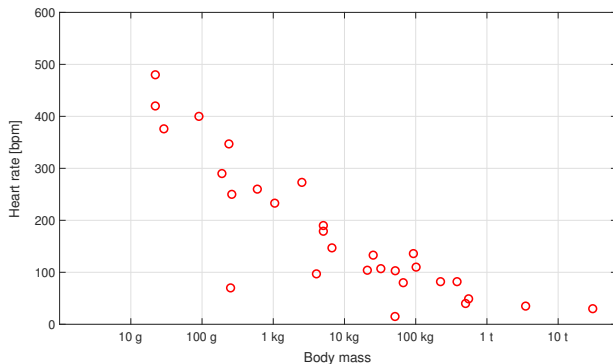
Dr Jesús Requena Carrión

28 Sept 2021

From mouse to whale



From mouse to whale, through rabbit



A rabbit's resting heart beats at

(a) ≤ 100 bpm

(b) ≥ 300 bpm

(c) ≥ 100 bpm and ≤ 300 bpm

Agenda

What is machine learning?

The value of knowledge

The machine learning taxonomy

About ECS7020P

Machine learning or statistical learning?

Machine or statistical learning is usually defined as

*The ability to acquire **knowledge**, by extracting patterns from raw **data**.*
(Goodfellow, Bengio, Courville)

*A set of tools for **modeling** and **understanding** complex **datasets**.*
(James, Witten, Hastie, Tibshirani)

What is data?

- **Data** is the materialisation of an **observation** or a **measurement**.
- **Datasets** are data formatted as collections of **items** described by a set of pre-defined **attributes**.

Animal (ID)	Body mass [g]	Heart rate [bpm]
Wild mouse	22	480
Rabbit	2.5×10^3	250
Humpback whale	30×10^6	30
...

In machine learning, **our data is always represented as a dataset**.

Note: Unfortunately, authors use different terms for the same concept. Item, sample, example, instance and point have the same meaning, and so do feature, variable and attribute. You should get used to all of them.

What is knowledge?

Knowledge can be represented as a

- **Proposition** (statement, law)
Smaller animals have a faster heartbeat.
- **Narrative** (description, story)
The size of an animal seems to be related to its heartbeat. In general, larger animals tend to have a slow heartbeat. For instance, the humpback whale...
- **Model** (mathematical or computer)
 $r = 235 \times m^{-1/4}$, where r is the heart rate and m is the body mass.

We will mostly use models to represent knowledge.

Knowledge as a model

Models describe **relationships** between attributes. **Mathematical** and **computer** models are equivalent:

- Mathematical models can be implemented as computer programs.
- Every computer model has a corresponding mathematical expression.

The mathematical expression $y = x + 3x^2$ is equivalent to the following Matlab line of code:

```
y = x + 3*x^2
```

or this Python line of code:

```
y = x + 3*x**2
```


Machine learning is **data science**...

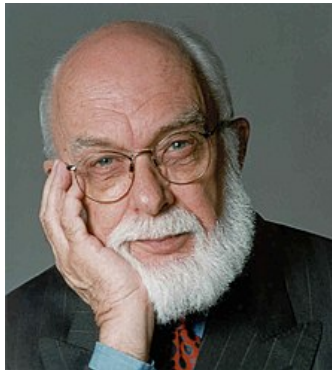
Science is not about using sophisticated instrumentation, maths or techniques, science is about **evaluating** our knowledge.

We use **data** together with **accepted knowledge** in this evaluation.

Proposition 1: *The earth is flat*

Proposition 2: *The earth is roughly spherical*

Pseudoscience



James Randi exposes dowsing:

www.youtube.com/watch?v=cqoYrSd94kA

*My concern is not **how** they do it, but **if** they do it.*

Data science is much **more than data**

There is no such thing as neutral data and raw data won't prove a model is true.

Craniometry (19th century): *The size of a brain is related to its degree of intelligence, e.g. big heads are smarter than small ones, or elongated heads are smarter than short ones.*

What about AI?

Definitions of AI include creating machines that act like humans, think like humans, act rationally or think rationally.

Some AI solutions use machine learning algorithms, some others do not. In addition, machine learning can be used outside the AI remit.

Most of the time, when media and companies talk about AI, they mean machine learning, basic statistics or just some form of computation.

We will approach machine learning rigorously. You might need to **forget** things that you have learnt!

Agenda

What is machine learning?

The value of knowledge

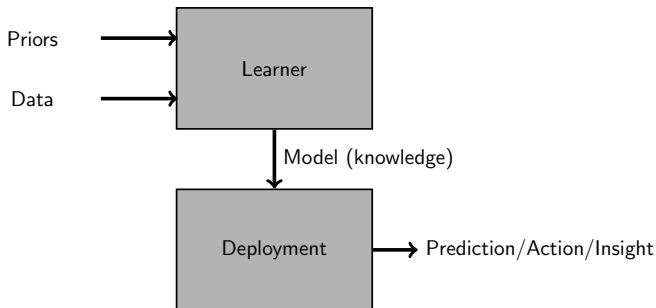
The machine learning taxonomy

About ECS7020P

The two stages of Machine Learning

Models can be built, sold and deployed to deliver **value**. During the life of a model, we can distinguish two stages:

1. **Learning** stage: The model is built.
2. **Deployment** stage: The model is used.



Deployment: eCommerce

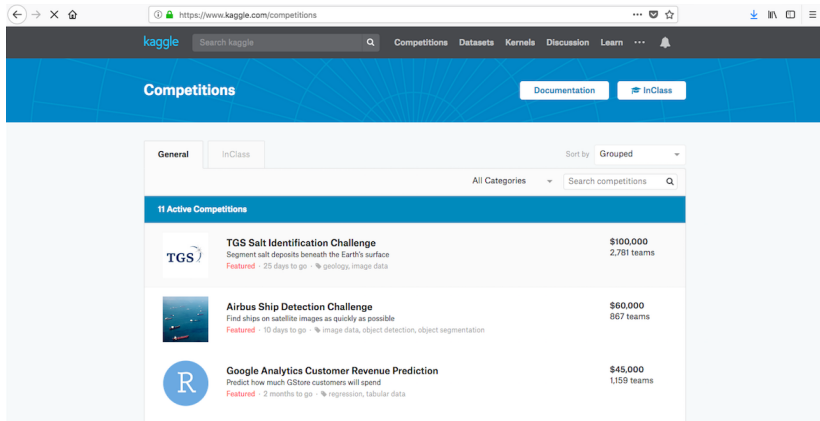
Inspired by your shopping trends






Recommendations for you in Grocery



Data science competitions



The screenshot shows the Kaggle website's 'Competitions' section. The browser address bar displays 'https://www.kaggle.com/competitions'. The page has a dark blue header with the Kaggle logo, a search bar, and navigation links for Competitions, Datasets, Kernels, Discussion, Learn, and a bell icon. Below the header is a blue banner with the word 'Competitions' and buttons for 'Documentation' and 'InClass'. The main content area has tabs for 'General' and 'InClass', with 'General' selected. A 'Sort by' dropdown is set to 'Grouped'. A search bar for competitions is present. A blue bar indicates '11 Active Competitions'. Three competitions are listed:

Competition Logo	Competition Name	Prize	Teams	Details
	TGS Salt Identification Challenge Segment salt deposits beneath the Earth's surface <i>Featured</i> · 25 days to go · 📁 geology, image data	\$100,000	2,781 teams	
	Airbus Ship Detection Challenge Find ships on satellite images as quickly as possible <i>Featured</i> · 10 days to go · 📁 image data, object detection, object segmentation	\$60,000	867 teams	
	Google Analytics Customer Revenue Prediction Predict how much GStore customers will spend <i>Featured</i> · 2 months to go · 📁 regression, tabular data	\$45,000	1,159 teams	

Machine learning basic methodology

In machine learning we are interested in finding **the best model**. Hence, we need a notion of **model quality**. However, our goal is to build models that work well **during deployment**, i.e. when presented with new data.

Basic machine learning methodologies include two separate tasks:

- **Training:** A model is created using data and a quality metric. We also say that we **fit a model** to a dataset.
- **Testing:** The performance of the model during deployment is assessed using new, **unseen data**.

Without rigorous methodologies, models are very likely to be of little use.

Agenda

What is machine learning?

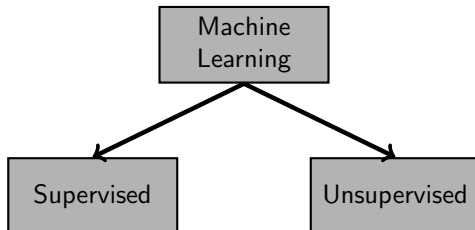
The value of knowledge

The machine learning taxonomy

About ECS7020P

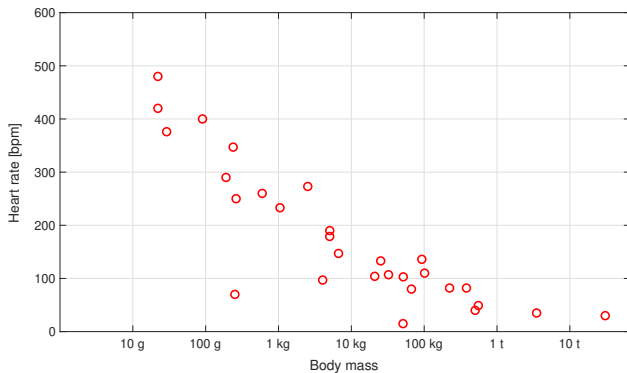
Problem formulation

What **types of problems** can we formulate in machine learning?



Supervised learning: Heart rate in the zoo

Can I guess the heart rate of an animal whose body mass I know, by looking at the heart rate and body mass of other animals?

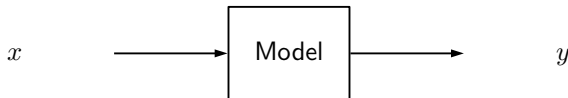


Supervised learning

In supervised learning, we are given a **new item** (*rabbit*) such that the value of one of its attributes is **unknown** to us (*rabbit's heartbeat*).

Our goal is to **estimate** (*guess*) the **missing value** by learning from a **collection of known items** (*weight and heart rate of other animals*).

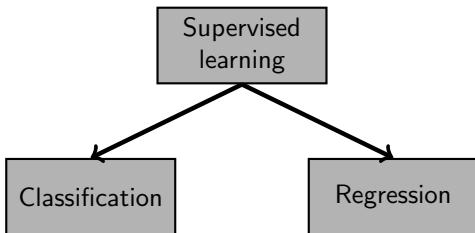
The challenge is then to build a model that maps one attribute x , known as the **predictor**, to another attribute y , which we call the **label**, using a dataset of **labelled examples**.



Supervised learning: Classification and regression

Supervised learning is further divided into two categories depending on the type of label:

- **Classification:** The label is a **discrete** variable.
In a spam detector, 0 could mean email is spam, 1 it isn't
- **Regression:** The label is a **continuous** variable.
The heart rate of an animal is a continuous label

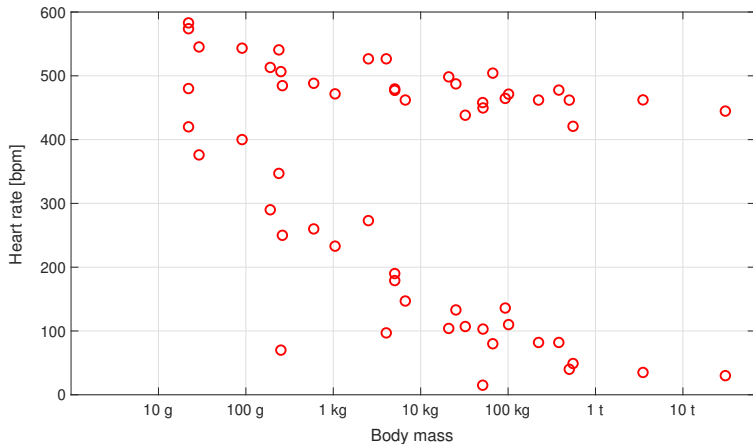


Unsupervised learning: Heart rate in the galactic zoo



Unsupervised learning: Heart rate in the galactic zoo

What can you conclude from this distribution of data points?



Unsupervised learning

In unsupervised learning, we set out to **find the underlying structure** of our dataset. This can be useful to gain understanding, identify anomalies, compress our data and reduce processing time.

Applications of unsupervised learning include:

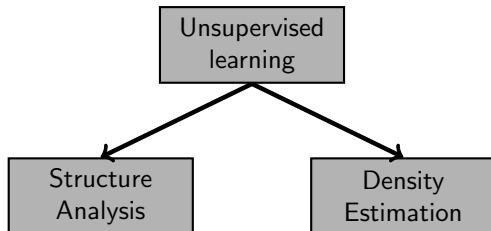
- Customer segmentation.
- Social community detection.
- Recommendation systems.
- Evolutionary analysis.
- Fraud detection.

Unsupervised learning

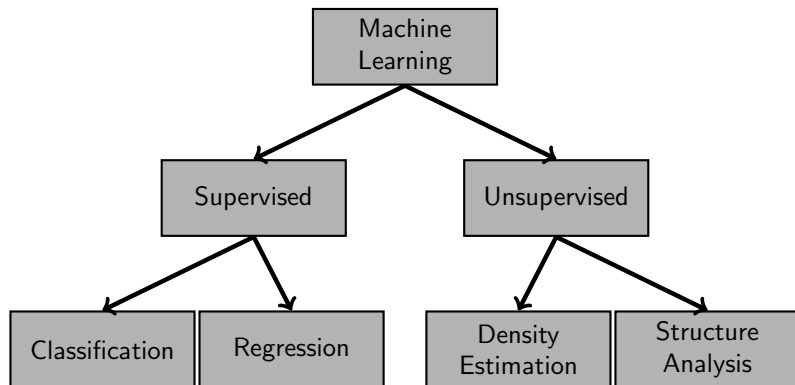
The underlying structure of a dataset can be studied using **structure analysis**, which includes:

- **Cluster analysis**: Focuses on groups of data points.
- **Component analysis**: Identifies directions of interest.

Density estimation techniques provide statistical models that describe the distribution of samples in the attribute space.



Machine Learning taxonomy



Agenda

What is machine learning?

The value of knowledge

The machine learning taxonomy

About ECS7020P

Learning goals

1. Understand the principles of ML, its scope and applications.
2. Be able to use the ML taxonomy to formulate meaningful questions and identify suitable techniques to answer them.
3. Discuss the relative merits of different ML techniques.
4. Be able to apply the methodology needed to build and evaluate ML solutions.
5. Be able to independently learn and confidently apply new ML techniques.
6. Critically analyse reports on ML applications and advances.
7. Understand model deployment and its main challenges.

Module contents

- Week 1: Introduction
- Week 2: Regression
- Week 3: Methodology I
- Week 4: Classification I
- Week 5: Classification II
- Week 6: Methodology II
- Week 7: Reading week
- Week 8: Neural networks & deep learning
- Week 9: Structure analysis
- Week 10: Density estimation
- Week 11: Deployment
- Week 12: Final revision

Study outline

This module consists of 150 study hours (lectures, labs, assessment preparation, study time, etc). Its duration is 12 weeks (week 7 is a revision week):

- **Lectures** (2h/week) will be delivered online using Blackboard Collaborate (QM+). MS Teams will be used as a fallback.
- **Tutorials** (1h/week) will take place online using Blackboard Collaborate (QM+). MS Teams will be used as a fallback.
- **Labs** (2h/week) will take place on-campus (ITL and End buildings) and online (MS Teams).

Check QM+ for more details.

Communication

- **Forum on QM+**. Primary means, questions might have been answered already and answers might be useful to others
- **Email**: Please make sure its subject is formatted as follows:
"[ECS7020P] <DESCRIPTIVE SUBJECT HERE>"
- **Face to face**: Physically or remotely (via MS teams).

Assessment and labs

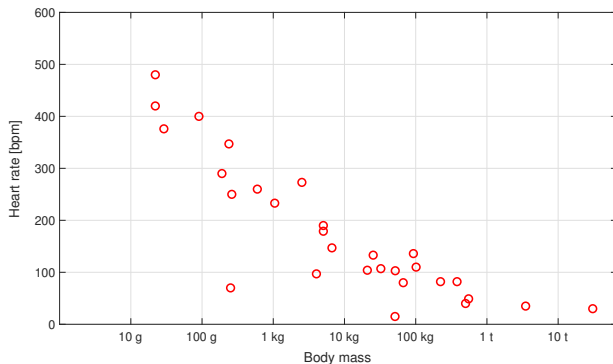
ECS7020P is assessed as follows:

- Final exam: 60 %
- CW: 40 %

CW activities use Python (Google Colab environment) and consist of:

- Lab-based quizzes (20 %). Weeks 1-6.
- Mini-project (20 %). We will create a dataset (weeks 3-6) and build a machine learning solution (weeks 8-11).

The strange case of the flatworm



The heart rate of a flatworm weighting less than 10 g

(a) Can't be guessed from this dataset

(b) Is ≥ 300 bpm

(c) None of the above

Know thy domain!