MTH6101: Introduction to Machine Learning
Semester B, 2023-24

Week 1, introductory slides

Kostas Papafitsoros & Hugo Maruri-Aguilar

Timetable

Lectures:

- (1) Tuesdays, 14:00-16:00 (Weeks 1-6, 8-12), Peoples Palace: Skeel-LT
  (2) Wednesdays, 09:00-11:00 (Weeks 1-6, 8-12), iQ East C. (Scape): 0.14
- (1) Thursdays, 14:00-15:00 (Weeks 1-6, 8-12), Gr. Ctr: G.10 (Peston LT)
  (2) Thursdays, 15:00-16:00 (Weeks 1-6, 8-12), Maths: MLT

IT-labs:

- (i) Fridays: 16:00-17:00 (Weeks 3-9, 11-12), Queens: QB-202
- (ii) Fridays: 17:00-18:00 (Weeks 3-9, 11-12), Bancroft: 1.15A
- (iii) Fridays: 17:00-18:00 (Weeks 3-9, 11-12), Bancroft: 1.23

!!! Only Week 2, Monday 29/01, Queens: QB-202:
11:00-12:00, 12:00-13:00, 13:00-14:00

Prerequisites

- Linear Algebra I and Calculus
- Statistical Modelling I (*essential*)
- Statistical Modelling II (*helpful*)
- Probability and Statistics

Programming language:

- **Rstudio**
- `https://cran.r-project.org`
- `https://rstudio.com/products/rstudio/download/`
- `https://rdrr.io/snippets`

→ Check instructions on Week 2 on QMplus!

Syllabus

Quick revision, Week 1:
Linear algebra, Calculus, Statistics

Unsupervised learning:
Weeks 2,3,4,5,6

(1) Principal Component Analysis (PCA)

(2) Data Clustering

Supervised learning:
Weeks 8,9,10,11,12

(1) Classification methods

(2) Penalised Likelihood

Syllabus

Quick revision, Week 1:
Linear algebra, Calculus, Statistics

Unsupervised learning:
Weeks 2,3,4,5,6

(1) Principal Component Analysis (PCA)

(2) Data Clustering

Mid-term Quiz, online
(15% of final mark)

Supervised learning:
Weeks 8,9,10,11,12

(1) Classification methods

(2) Penalised Likelihood

End-term Quiz, online
(15% of final mark)

Assessment

Mid-term Quiz, online (15% of final mark)

Week 7, Friday, 10-12am

End-term Quiz, online (15% of final mark)

Week 12, TBA

Final exam, online (70% of final mark)

Exam period, TBA

Assessment

Mid-term Quiz, online (15% of final mark)

(You will have access to the correct answers after the end of the quiz)

Week 7, Friday, 10-12am

End-term Quiz, online (15% of final mark)

(You will have access to the correct answers after the end of the quiz)

Week 12, TBA

Final exam, online (70% of final mark)

(Some previous exam questions with their solutions already available in QMplus)

Exam period, TBA

Communication/Questions/Discussions

→ Office hours: Mondays 14:00-16:00, MB117 (or via teams)
→ You can you ask questions on teams and the student forum
→ Learning Cafe (more on that later)

Communication/Questions/Discussions

→ Office hours: Mondays 14:00-16:00, MB117 (or via teams)
→ You can you ask questions on teams and the student forum
→ Learning Cafe (more on that later)

There is no such thing as a stupid question

Communication/Questions/Discussions

→ Office hours: Mondays 14:00-16:00, MB117 (or via teams)
→ You can you ask questions on teams and the student forum
→ Learning Cafe (more on that later)

There is no such thing as a stupid question

(and we are very approachable)

What is machine learning?

Machine learning:

Umbrella term for a collection of **automated methods** methods for data analysis.

Machine learning:

Umbrella term for a collection of **automated methods** methods for data analysis.

The target is typically to describe future behaviours (**predict**) that will help perform some decision making.

- *You would like to have a tool that will tell you (predict) which banking transactions are fraudulent and which not*
- *You would like to know which customers exhibit similar travel behaviour to each other*

Machine learning:

Umbrella term for a collection of **automated methods** methods for data analysis.

The target is typically to describe future behaviours (**predict**) that will help perform some decision making.

- *You would like to have a tool that will tell you (predict) which banking transactions are fraudulent and which not*
- *You would like to know which customers exhibit similar travel behaviour to each other*

What does **"automated methods"** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to "traditional" approaches?

What does **“automated methods”** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to “traditional” approaches?

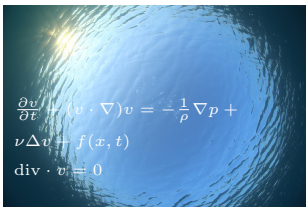
What does **"automated methods"** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to "traditional" approaches?

How was science traditionally done:

What does **“automated methods”** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to “traditional” approaches?

How was science traditionally done:

Build a model



Collect data

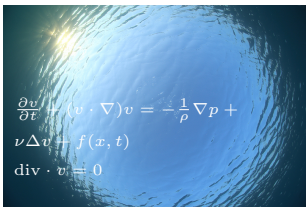
www.marinedataservice.com, CC BY-SA 4.0

Test the model

What does **“automated methods”** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to “traditional” approaches?

How was science traditionally done:

Build a model



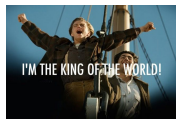
Collect data

www.marinedataservice.com, CC BY-SA 4.0

Test the model

Model agrees with the data

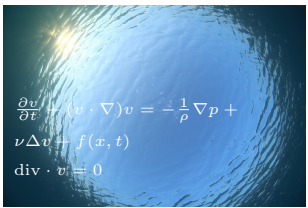
✓



What does **“automated methods”** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to “traditional” approaches?

How was science traditionally done:

Build a model



Collect data

www.marinedataservice.com, CC BY-SA 4.0

Test the model

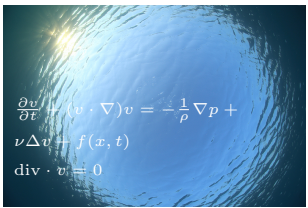
Model disagrees with the data

×

What does **“automated methods”** mean here?
What is the distinctive characteristic of machine learning?
What is the difference to “traditional” approaches?

How was science traditionally done:

Build a model



Collect data

www.marinedataservice.com, CC BY-SA 4.0

Test the model

Model disagrees with the data

×

Build a better model

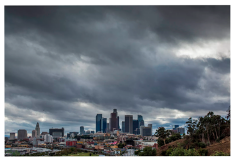
In machine learning:

- $\Rightarrow$ The use of **available data** is an essential ingredient towards solution of the task and decision making
- $\Rightarrow$ Detect patterns in **available data** and then use these patterns to make predictions for other similar data

The model is built ("learned") from data!

Google says new AI models allow for 'nearly instantaneous' weather forecasts

An increasingly important tool in a world shaped by climate change



What changed? (from early 2000's)

- *Huge amount of data* became gradually available
- *Increased computing power* to deal with so much data
- *Development of sophisticated algorithms*

Examples:

- *You would like to have a tool that will tell you (predict) which banking transactions are fraudulent and which not*

What changed? (from early 2000's)

- *Huge amount of data* became gradually available
- *Increased computing power* to deal with so much data
- *Development of sophisticated algorithms*

Examples:

- *You would like to have a tool that will tell you (predict) which banking transactions are fraudulent and which not*

  $\rightarrow$ You have now a lot of available data of already made transactions which **you know** that they are fraudulent
  $\rightarrow$ Use them, in order to find out what patterns characterize fraudulent transactions

What changed? (from early 2000's)

- *Huge amount of data* became gradually available
- *Increased computing power* to deal with so much data
- *Development of sophisticated algorithms*

Examples:

- *You would like to know which customers exhibit similar travel behaviour to each other*

  $\rightarrow$ You have now a lot of available data of customers and their travel behaviour
  $\rightarrow$ Use a *clustering technique* to group them together

There are roughly two types of methods in machine learning:

Supervised learning

Unsupervised learning

There are roughly two types of methods in machine learning:

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Unsupervised learning

There are roughly two types of methods in machine learning:

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

→ Predict labels for new data or an output given some new input

Unsupervised learning

- Available data do not come with known labels
- There are no input-output data pairs

→ Learn patterns, identify distinct groups, generate new examples

Examples

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Examples

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Classification: Assign data to categories (labels)

£££	£££	£££	⋯	£££	£££
Transaction	Transaction	Transaction		Transaction	New transaction
£££	£££	£££		£££	£££
↓	↓	↓		↓	↓
Fraudulent	Normal	Fraudulent		Normal	???

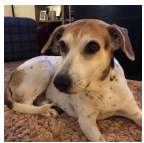
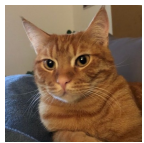
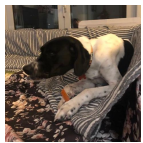
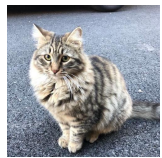
Examples

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Classification: Assign data to categories (labels)




...



↓ Cat ↓ Dog ↓ Cat ↓ Dog ↓ ???

Examples

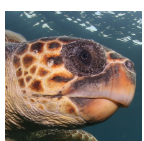
Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Classification: Assign data to categories (labels)




...

↓ Turtle A

↓ Turtle B

↓ Turtle A

↓ Turtle B

↓ ???

Examples

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Classification: Assign data to categories (labels)




...

↓ Turtle A ↓ Turtle B ↓ Turtle A ↓ Turtle B ↓ ???

Methods: **Logistic regression**, **decision trees**, **linear discriminant analysis**, **K-nearest neighbour**, support vector machines, neural networks...

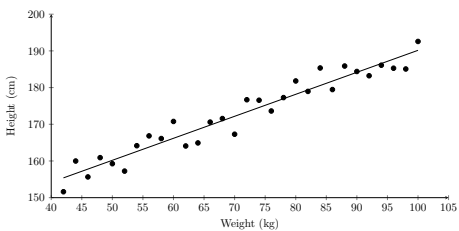
Examples

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Regression: Predict continuous quantities from input data



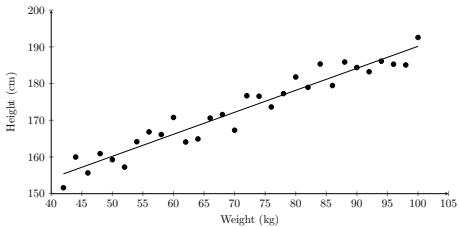
Examples

Supervised learning

- Available data come with known labels
- Input-outputs pairs are specified for these data

$\rightarrow$ Predict labels for new data or an output given some new input

Regression: Predict continuous quantities from input data



Methods: **Linear regression**, nonlinear regression, Gaussian processes, neural networks...

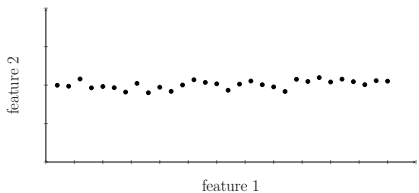
Examples

Unupervised learning

- Available data do not come with known labels
- There are no input-output data pairs

$\rightarrow$ Learn patterns, identify distinct groups, generate new examples

Dimensionality reduction: Map high dimensional data into low dimensions while still keeping relevant information



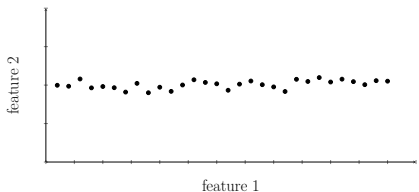
Examples

Unupervised learning

- Available data do not come with known labels
- There are no input-output data pairs

$\rightarrow$ Learn patterns, identify distinct groups, generate new examples

Dimensionality reduction: Map high dimensional data into low dimensions while still keeping relevant information



Methods: **Principal component analysis**, factor analysis, neural networks...

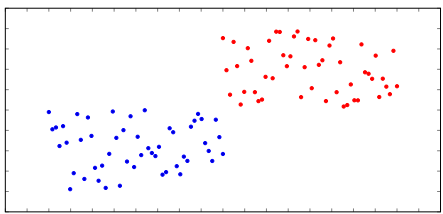
Examples

Unupervised learning

- Available data do not come with known labels
- There are no input-output data pairs

$\rightarrow$ Learn patterns, identify distinct groups, generate new examples

Clustering: Organise data in groups of similar points



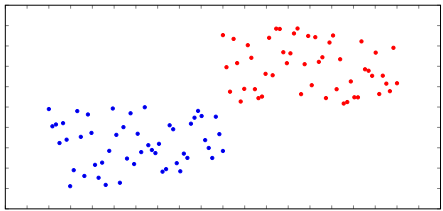
Examples

Unupervised learning

- Available data do not come with known labels
- There are no input-output data pairs

$\rightarrow$ Learn patterns, identify distinct groups, generate new examples

Clustering: Organise data in groups of similar points



Methods: **Agglomerative clustering**, **K-means clustering**

Summarizing:

Supervised learning

- **Classification:** To which category does this data point belong?
- **Regression:** Given this input from a data set, what is the likely value of a particular quantity?

Unupervised learning

- **Dimensionality reduction:** What are the most significant features of the data and how can they be summarised/visualised?
- **Clustering:** Which data points are similar to each other?

Other types (not covered in this module):
Neural networks, deep learning, semi-supervised learning, reinforcement learning...