Applied Machine Learning - Class 1

Professor Michael Dowling, Dublin City University

A brief overview of the module

About me

Professor of Finance in Dublin City University.

Previously worked in Rennes School of Business, where I was Director of the AI-driven Business research centre.

A lot of published research on applied machine learning, primarily concentrating at the moment on machine learning of textual data.

Course delivery

23/05/2022	11:00 - 13:00
06/06/2022	11:00 - 13:00
20/06/2022	11:00 - 13:00
20/06/2022	17:00 - 19:00
04/07/2022	11:00 - 13:00
04/07/2022	17:00 - 19:00

Purpose of the module

You have all done a Python data science class before, and this module deepens that knowledge by giving you a machine learning and deep learning skillset.

We are primarily, as the module title suggests, concerned with **applied** machine learning knowledge. I'll cover the rationale for the methods with you, but mainly we will be **doing**.

Module structure

1. Today we concentrate on a practical overview of all the main methods, as well as explaining the general idea of machine learning
2. For the rest of the course we will deepen this practical knowledge, backed up by recommended additional learning and practical tasks
3. This leads to the assessment of the module, which will be a high-quality delivered group project by you on applied machine learning

Assessment

The module is 100% assessed by a group project (groups of four), following standard Bayes group project rules.

You will be asked to analyse a dataset through applying a range of machine learning and deep learning techniques, and then produce an analytic report showing strong practical implications from your investigation.

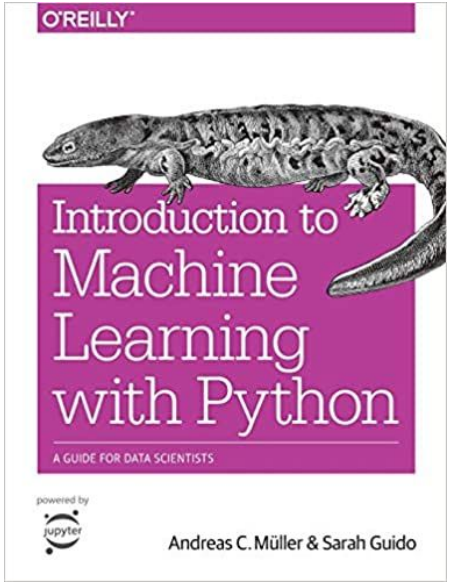
Deadline for submission is Monday 18th July.

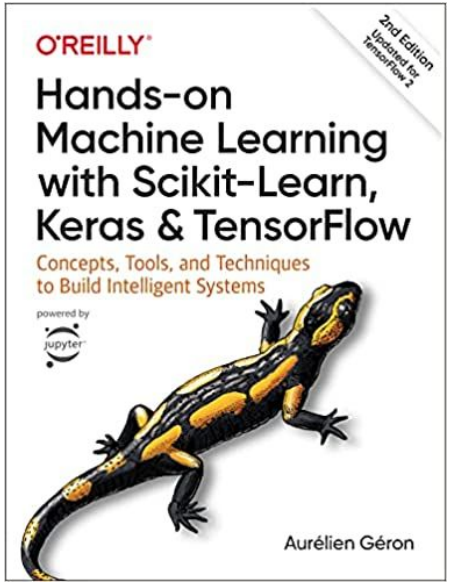
Reading and further information

Have a look at the module structure on Moodle for more detailed information on the course.

There are two good textbooks for the course, both of which are handy as a future reference library when you want to apply these techniques again.

First textbook is very good, very clear. Second textbook is more ‘industry-proof’ in terms of advanced knowledge. Quite a bit of overlap between the two books.





Today’s topics

Topic breakdown

1. What is machine learning
2. The popular tools of machine learning
3. Applied overview of machine learning
    a. Data wrangling
    b. Traditional testing
    c. Supervised machine learning
4. Tasks for the next class

What is machine learning?

What is machine learning?

Machine learning (ML), like data science in general, is about extracting knowledge from data.

Often confused *by the layperson* with artificial intelligence (AI), however AI involves the addition of some form of automated decision making.

A contrast is with traditional regression, where the tester pre-determines expected relationships between variables.

ML *sometimes* start with this same 'expected relationship' basis (known as **supervised learning**), and sometimes start with almost no theory (known as **unsupervised learning**).

However ML always goes beyond traditional testing, by allowing the *machine*, or algorithm, to iterate to its own learning outcomes.

Some casual watching in your own time - the promise of ML

And a brief clip with the same idea

Note: this is indeed AI, rather than just ML, as the algorithm has used ML to iterate to a good learning outcome, and then implemented the learning itself.

Supervised learning

Supervised learning is the most common form of ML and starts with a dataset that has a good spread of example **input-outcome** data spread across all feasible outcomes:

- **if** email contains BuY ViAgRa ChEaP **then** email is spam
- **if** email from regular correspondent **then** email is not spam

With sufficient data examples, supervised learning can perform excellently at identifying relationships between input data and outcomes, and these models stand up well to **predicting** unknown outcomes from new input data.

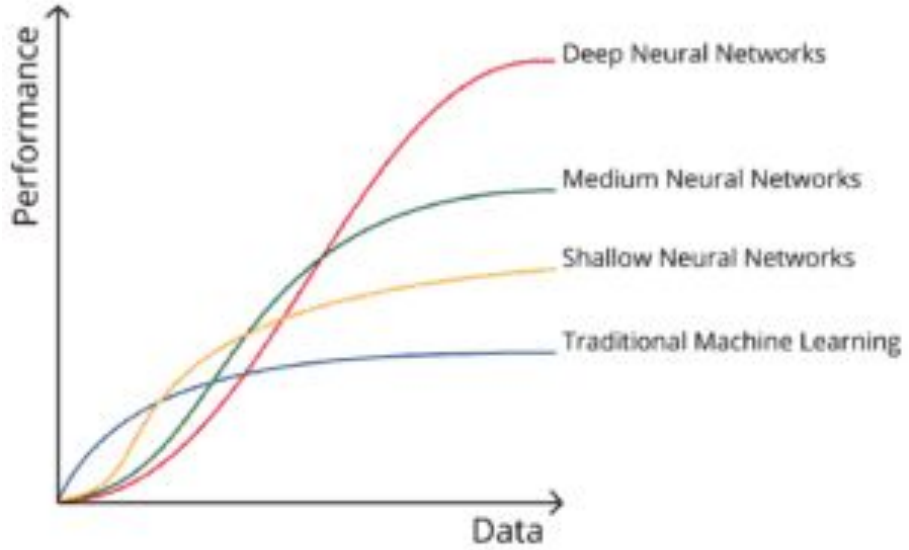
Examples of supervised learning

- Predicting outcomes in general as an improved version of traditional regression techniques
- Identifying groups within data
- Uncovering decision paths that lead to certain outcomes
- Identifying outlier outcomes

More practically:

- Identify determinants of purchase decisions, identifying financial market and other market pricing relationships, detecting fraudulent and other outlier outcomes

The importance of ‘more’ data



A note on machine learning vs deep learning

Geoffrey Hinton is often credited with ‘the modern deep learning revolution’ for his work since 1986. Although, *as he himself frequently emphasises*, the revolution was created by teams of people rather than a single individual.

The broad insight of deep learning is that, whereas ML attempts to find a direct relationship between input data and outcomes, deep learning recognises that there are probably intermediate layers (**hidden layers**) through which data passes before an outcome is decided. This process is viewed as mimicking to some extent how the brain makes decisions.

Deep learning integrates these hidden layers in its modelling.

While this has been proposed for decades, the particular contribution of Hinton and others was to identify a new algorithmic approach that appears to allow better learning from these layers.

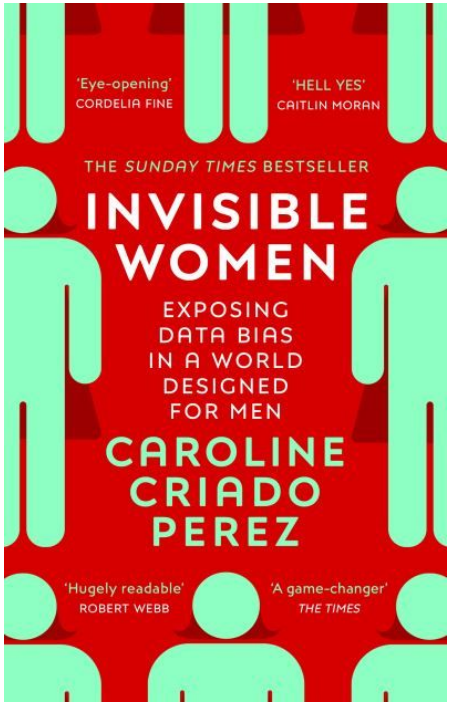
Unsupervised learning

Unsupervised learning is the second core category of ML. The key difference from supervised learning is that only the input data is known and we are trying to identify feasible outcomes.

Many of the same techniques in supervised learning can also be used as unsupervised learning, with adjustments. For example, the techniques of deep learning can be either supervised or unsupervised.

We'll talk about this category a bit more later in the course.

A note on bias in machine learning



Uncontrolled ML has been linked to multiple examples of bias, including gender and racial bias. This is frequently unintended bias, **yet unintended bias is still bias**. We need to carefully consider our input data for any aspects that can create bias.

There's an excellent analysis of this tendency here: https://www.brookings.edu/research/fairness-in-algorithmic-decision-making/ and a great data-driven analysis in the book 'Invisible Women' by Caroline Criado Perez

The popular tools of machine learning

Python

Python, and its packages, is the default tool of machine learning.

R can also be used, but it seems like a poor idea to choose a programme that is far less popular.

Have read of this quite well-balanced article on Python vs R in finance: https://www.efinancialcareers.com/news/2021/09/banks-python-vs-r - essentially, even if R has its advantages, **Python is what people are actually using**.

Which Python packages

NumPy and **pandas** are the core data analytic packages

scikit-learn (also known as **sklearn**) is the core machine learning package

Apart from that you need some sort of charting software (**matplotlib** as the most basic iteration) and then some way to access Python.

We'll discuss some other packages as we go through the course.

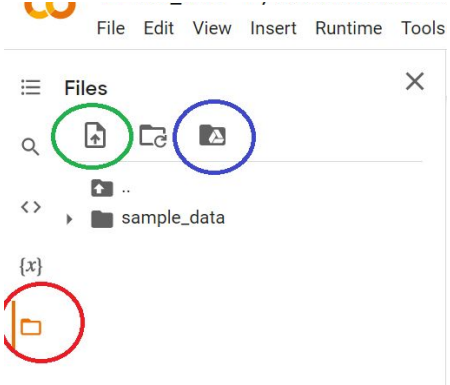
Accessing Python

I know you already know this, but I want to cover the range of options for accessing Python here:

1. **colab.research.google.com** is the most basic way. All core packages pre-installed, hosted on a cloud. Some testing limits, but these can be removed by upgrading to premium.
2. **Jupyter Notebooks** allows you to run Python through your own computers power. https://jupyter.org/.
3. **PyCharm** (or equivalent) as a popular full professional coding solution.
4. Hosting a Python notebook through **Google Cloud** or Amazon / Microsoft - a task for our next class.

Applied overview of machine learning

Steps to follow



Go to **colab.research.google.com** logging in with your google account. Feel free to use a Jupyter Notebook if you have it pre-installed from an earlier class, but we won't have time during the class to address major issues with Jupyter Notebook

File > Upload notebook > upload the notebook called **Class1.ipynb** (get it from Moodle)

In the sidebar on Colab (see image) upload the dataset **Class1.csv** (get it from Moodle), or load the file from your Google Drive.

Let's go!

Tasks for the next class

Things to do to get ahead

Read and practice a little:

Read the **first chapter** of the *Introduction to Machine Learning with Python* textbook. Its very nicely written, it won't cause you to freak out about the course, its short.

Try implementing the code for the **iris dataset** example given in the chapter.

Deepen your applied workplace knowledge

Also on Moodle, I have included some additional readings. These are intended to familiarise you with some practical aspects of how machine learning might be implemented in the workplace:

1. Read the document created by the previous person who delivered this course, Alan Chalk, on **reproducibility**. Reproducibility of your models is vital in the workplace for consistency, and often for regulatory reasons.
2. h2o.ai is a popular **automated** ML tool. It would be worth signing up to the 90-day free trial and exploring what automated ML looks like. There's some excellent tutorials on there.