IMPERIAL

Department of Mathematics

MATH60026/MATH70026
Methods for Data Science
Years 3/4/5

Academic year 2024-25

Spring term
Course overview

People

Lecturer:
Dr Barbara Bravi

Graduate Teaching Assistants:
Zekai Li
Kevin Michalewicz
Ella Orme
Dominik Schindler

Schedule

The course runs from 13/01/2025 (first lecture) to 17/03/2025 (last lecture).
We have:

2 hours on **Monday** (4-6pm, HXLY 340, HXLY 130 as overflow room): **Lecture**
1 hour on **Tuesday** (1-2pm, HXLY 213): **Tutorial on notebooks + Q&A**

The material (lecture notes + coding notebooks) is released in advance on Blackboard, so that you can start studying the week's topic, test your understanding and develop your algorithmic skills through the coding notebooks.

Specifically:
Thursday of the week before at 12pm: release of lecture notes and unsolved notebook;
Monday at 4pm (just before the lecture) = release of the lecture slides;
Tuesday at 1pm (just before tutorial) = release of the notebook solutions.

The lecture notes contain a few instructive but long mathematical derivations that I won't not go through in detail in lecture; rather I will explain the key results and build upon them to discuss method implementations and applications to data analysis.

For more questions: **office hour** after the tutorial (**Tuesday, 2-3pm**) in HXLY 410.

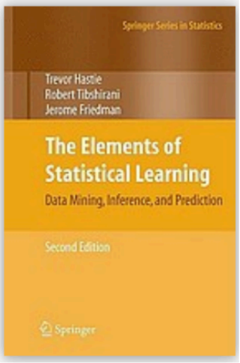
Course materials

There will be a signposting document to guide you through the materials of each week.

They will include:

- **Lecture Notes:** the lecture notes cover the topics you need to learn for the course + some extra topics for your own interest (for which we provide additional references).

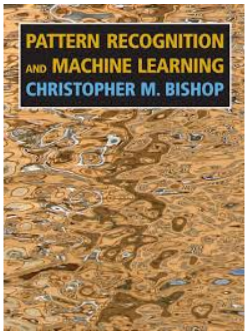
The notes also provide pointers the **textbooks** recommended for the course:


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Main textbook:

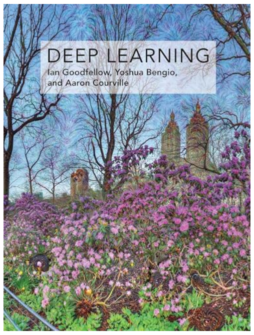
The Elements of Statistical Learning: Data Mining, Inference and Prediction

by Hastie, Tibshirani, Friedman



Deep Learning

by Goodfellow, Bengio, Courville



Pattern Recognition & Machine Learning

by Bishop

We upload the slides of the lecture on Blackboard, along with some extra-reading material (like research papers on methods studied) in the folder ‘Additional reading’.

- **Notebooks with Coding tasks:** coding notebooks with brief introductions and explanations. The notebooks contain important algorithms (incomplete with hints) to be completed as your coding tasks. Solutions for the coding tasks will be provided every week and discussed during the tutorial session on Tuesdays.

Remark: Lecture notes and notebooks are provided for the personal study and training of students of the course. The distribution of copies in part or whole is not permitted.

Coding language

We use **Python** (Jupyter notebooks) for computational tasks and for the Coursework, due to its wide use in Data Science and Machine Learning.

Among the different options, you can generate Jupyter Notebooks on your machine using VS code (see e.g. here). Alternatively, you can use Google Colab Notebooks.

This is not a coding course, but reasonably good coding is a necessary skill to develop. There is complete online documentation on Python: part of your skill development and workload will be working on your own through this material.
The computational tasks released weekly associated with the material covered in the lectures will help you develop the necessary coding skills.

Assessment

The assessment is based on **coursework**. There will be **two assessed courseworks** (CW1 and CW2).

Each CW is like a **data science mini-project:** it involves the analysis of data sets through the algorithms taught in class, using the code developed in the coding tasks as your building blocks. 4th year students and MSc students will have a mastery component to perform, alternative to some tasks of the BSc coursework.

Some questions will be a bit more open-ended and aimed at testing your understanding; high-quality presentation (plots, explanations) is expected.

Examples of previous courseworks and some additional guidelines on presentation quality will be made available on Blackboard.

Structure of the courseworks:

- Coursework 1 (weight: 40% of the mark)

You will have 2 weeks to work on CW1, which is:

Released: Friday, **7 February 2025 at 1pm**

Due: Friday **21 February 2025 at 1pm**

You will commit to the course by handing in CW1.

- Coursework 2 (total weight: 60%)

CW2 is divided into two parts:

1. Part 1 of CW2 (30% of CW2): similar to CW1, you have 1 week to work on it.

Released: Wednesday, **12 March 2025 at 1pm**

Due: Wednesday, **19 March 2025 at 1pm**

2. Part 2 of CW2 (70% of CW2): a series of tasks to be performed in a controlled environment (in class) and in a limited time frame, like in a class test. You will work on the same data as in Part 1 of CW2, you will be given 2 hours.

   Date: Friday, **21 March 2025** (details on time and room to come)

   Important: we are adopting the Secure Exam Workspace (SEW), giving you access only to permitted software, browsers & materials (instructions to come).

   With the advent of generative AI software, the College guidelines on assessments require additional measures to ensure the authenticity of the work submitted by each student; SEW is the only truly safe option to ensure that the rules we have set to avoid plagiarism (see below) are respected. Invigilators will be present as well.

   SEW familiarisation session: Monday**, 10 March 2025 2-4pm** (4 sessions of 30 minutes each).

Logistics: You will produce Jupyter notebooks for both CWs, which will be handed in **via Blackboard** (we will provide detailed instructions).

Resits: They will be run **in August**, and for extenuating circumstances, another full CW2 (comprising both Part 1 and in-class Part 2) might be available.

Rules to avoid plagiarism: Needless to say, projects must be your own work.

You may discuss the analysis with your colleagues but the code, writing, figures and analysis *must be your own*. The Department uses code profiling and tools such as **Turnitin to check for plagiarism, and plagiarism is a form of misconduct that cannot be tolerated**. Cases of plagiarism have been detected in past years, with serious consequences on the students' academic results.

Note that submitting **AI-generated content is considered a form of plagiarism**, because it is work not created by you (see pages 33-34 of the Student Handbook 2024-2025). Hence **the use of generative AI (ChatGPT and similar software, like GitHub Copilot, Gemini, Bing etc) is not allowed** in the coursework you submit, because it is not functional to the learning objectives of the course. This course is an opportunity to learn skills (in terms of algorithm design as well as basic programming) that are fundamental for a data scientist.

Even if your future job relies on AI tools that facilitate and automate code writing, these skills are key to a clever, correct and effective use of these tools.

Ed Forum:

There will be a link within Blackboard to an **online student forum on EdStem.**

We will use EdStem as a **peer-to-peer forum**, where you can comment and help each other by providing hints or pointers to material for the tasks and lecture notes. You should not share pieces of code or answers for the courseworks on the forum.

The EdStem Forum is **not a 24/7 helpline.** The forum will be monitored both by the lecturer and the senior GTAs, to check that the forum is used correctly, and to identify potential issues and questions that emerge in the discussions. We will collect important questions, and cover some of these questions during the tutorial.

Please, use the lecture, tutorial and office hour for the questions for us.

To facilitate communication and feedback to us, we will appoint a class representative.

Tips on good practice on EdStem:

1. Before asking a question, please **check if the same question has been already asked** and answered.

2. **Keep your questions public** (unless they are related to a personal issue): even if your question is public (i.e., visible to other students), you can ask it **anonymously**. Keeping the questions public makes the discussion beneficial to all.