IMPERIAL

Department of Mathematics

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

Academic year 2024-25

Spring term
Course overview

Content and aims of the course

Aims: This is an introductory course that covers the **mathematical concepts underpinning** several of the most popular **methods used in learning from data**. The focus will be in understanding the mathematical concepts **through computational tasks**.

The course covers a very **broad collection of methods**. Therefore it remains at the introductory level focussing on concepts rather than detail. It will introduce mathematical formalism, and precise formulations for different heuristics, with few detailed proofs, but then the **focus will be on the implementation of algorithms** in their full mathematical structure.

The module will need a substantial amount of coding in **Python.**

The course will also cover some examples where the methods are used, and the coursework will try and **develop the skills to attack data science problems:**

1. inspecting and examining the data
2. posing mathematical and statistical questions to be answered
3. choosing the most appropriate methods
4. analysing the outcomes in a reasoned and critical manner
5. translating the results into clearly explained messages to be presented to non-experts.

Syllabus: In broad terms, the syllabus has three parts, of unequal length:

Part I. Supervised Learning (longer)
Part II. unsupervised learning
Part III. Graph-based Learning (shortest)

Part I: Supervised Learning. We will describe the mathematical ideas in learning from data the mapping between a certain input and an output, for instance the assignment to a class, or the prediction of an outcome. It will consist of theory and application of commonly used methods for the tasks of regression and classification (such as k-nearest neighbours, random forests, support vector machines, neural networks).

Week 1: Linear Regression
Week 2: k-Nearest Neighbours & Logistic Regression
Week 3: Naive Bayes & Random Forests
Week 4: Support Vector Machines
Week 5: Neural Networks - Multi-Layer Perceptron
Week 6: Convolutional Neural Networks

Part II: Unsupervised Learning. We will introduce unsupervised learning methods as tools to extract and inspect the intrinsic structure and properties of the data. This part will focus on the tasks of dimensionality reduction and clustering, including methods such as principal component analysis, non-negative matrix factorization, k-means, hierarchical clustering, clustering based on probabilistic mixture models.

Week 7: Clustering
Week 8: Dimensionality Reduction

Part III: Graph-based Learning (Guest lecturer: Dr Hardik Rajpal). We will illustrate with examples the power of graphs to model relationships in data. We will introduce key concepts of graph theory as relevant for data analysis (such as spectral graph theory and centrality measures); there will also be an introduction to some more advanced topics and applications.

Week 9: Graph-based learning
Week 10: Research seminars (not examinable)