MSc Machine Learning and Data Science Module Guide 2025-26

Module schedule

Year	Term	Module Code	Module Name	Type	ECTS
1	1	MATH70095	Applicable Maths	Core	5
1	1	MATH70094	Programming for Data Science	Core	5
1	2	MATH70098	Ethical Machine Learning and Data Science (Part 1-2) P1	Core	3.75
1	2	MATH70096	Exploratory Data Analytics and Visualisation	Compulsory	5
1	2	MATH70097	Supervised Learning	Compulsory	7.5
1	3	MATH70102	Unsupervised Learning	Compulsory	7.5
1	3	MATH70100	Bayesian Methods and Computation	Compulsory	7.5
2	4	MATH70104	Learning Agents	Compulsory	5
2	4	MATH70103	Unstructured Data Analysis	Compulsory	7.5
2	5	MATH70098	Ethical Machine Learning and Data Science (Part 1-2) P2	Core	3.75
2	5	MATH70101	Deep Learning	Compulsory	7.5
2	5	MATH70099	Big Data: Statistical scalability with PySpark	Compulsory	5
2	6	MATH70105	Research Project	Compulsory	20

The information displayed in this guide is correct at the time of publication and is subject to change.

Module Descriptions

Year 1 – Term 1

MATH70095: Applied Mathematics

Module Leader: Dr. Alexander Modell

Description

This module will provide you with the mathematical and statistical tools which will put you in the best position to succeed in your later modules on the MLDS programme. We will review the fundamentals of calculus, linear algebra, probability theory, statistics and optimisation.

Learning outcomes

- On successful completion of this module, students will be able to:
- Solve problems involving calculus in uni- and multivariate settings;
- Apply tools from linear algebra including matrices and their decompositions;
- Explain and use fundamental concepts in probability, including probability spaces, random variables and stochastic processes;
- Create and interpret summaries of data using sample-based statistics and explain some fundamental results in theoretical statistics;
- Understand how to measure distances and similarities between data, and implement various optimisation techniques;
- Apply relevant mathematical knowledge to address problems in machine learning

Module content

- Review of Calculus
- Linear Algebra
- Matrix Decomposition
- Introduction to Probability Theory
- Random Variables & Probability Distributions
- Concentration Bounds & Limit Theorems
- Stochastic Processes
- Sample-Based Statistics
- Optimisation
- Distance Functions

MATH70094: Programming for Data Science

Module Leader: Dr. Randolf Altmyer

Description

This module will provide the skills and knowledge to support the implementation, test and deployment of Machine Learning algorithms, and to construct data processing and analytic pipelines for Data Science. The module will compare and contrast R and Python, the two most popular Data Science languages, and students will become fluent using both languages for the range of challenges arising in practical data analysis problems.

Learning Outcome

- Fluently manipulate and transform data formats, using libraries where appropriate
- Select suitable data structures for different tasks and manipulate and transform such structures
- Mitigate problems with data and code using appropriate tools
- Refine code using appropriate tools
- Use R and Python fluently and appropriately

Module Content

- Data input and output
- Data structures
- Libraries
- Good programming practice, code and data
- Profiling
- Debugging R and Python

Year - 1 Term 2

MATH70098: Ethical Machine Learning and Data Science

Module Leader: Dr. Zak Varty

Description

This module will investigate the ethical implications of the new capabilities offered by Data Science and Machine Learning. Part 1 will begin by discussing the need for ethical considerations as part of a data science workflow through a discussion of real-world examples of failings and adverse outcomes. It will then introduce the sets of principles that tech leaders and international bodies are adopting to promote ethical use of data science and artificial intelligence algorithms. Privacy, fairness and accountability be explored in detail. Parts 2 will explore the remaining principles of ethical data science. We will consider in detail the explainability of ""white-box"" and ""black-box"" models, how to find patterns within data that generalise to new contexts and outline the safety vulnerabilities of distributed learning systems to adversarial attacks.

Learning Outcome

On successful completion of this module, students should be able to:

- Recognise and accept responsibility for the societal impact of data science and machine learning technologies; Participate in the broader debate about the issues surrounding the use of data science and machine learning for prediction, decision making and knowledge generation tasks;
- Identify common ethical pitfalls of data science and ML algorithms via a mental “check-list” and evaluate the degree to which a given algorithm is likely to conform with ethical best practices.
- Formally test for common ethical pitfalls of data science and ML algorithms. Implement mitigation measures against the ethical risks posed by the use of data science and ML algorithms.

- Construct well-founded and evidence-based arguments with which to positively influence the actions of stakeholders and decision-makers;
- Use a systems perspective to holistically appraise data science projects on their ethical and societal impacts.

Module Content

- Motivation and frameworks for ethical MLDS
- Measuring and ensuring privacy in data science workflows
- Measuring and ensuring fairness in data science workflows
- Interpretability for both ""white-box"" and ""black-box"" models.
- Causal inference and experimental design as tools for finding generalisable relationships
- Risks and benefits of learning from distributed data.

MATH70096 - Exploratory Data Analytics and Visualisation

Module Leader: Prof. Niall Adams

Description

This module will provide the skills and knowledge required to produce convincing narrative summaries and informative visualisations for a variety of complex datasets. The module covers assessing the structure and evaluating the quality of data and outlines techniques which uncover the underlying structure in data, both for initial reporting to a variety of intended audiences and to provide guidance for potential formal analysis and model formulation. The analysis and visualisations will be predominantly implemented using the suite of R tidyverse packages.

Learning Outcome

On successful completion of this module, students should be able to:

- Identify different data formats (categorical, continuous, discontinuous), in order to identify appropriate techniques;
- Collect data in formats ready for analysis
- Extract features of data to determine underlying structure
- Summarise/condense data in formats that make them interpretable by intended users
- Present visually compelling representations of the data in order to understand the structure for further modelling or for direct interpretation
- Produce convincing narratives for different audiences with an understanding of ethical considerations

Module Content

Data measurement scales Tidyverse and organisation of data frames Data quality issues, missing values and outliers Univariate and bivariate descriptive statistics Grammar of Graphics, ggplot2 Data Transformations Techniques for spatiotemporal data Techniques for multivariate data Techniques for unstructured data.

MATH70097 – Supervised Learning

Module Leader: Dr Nicola Gnecco

Description

This module introduces the framework of supervised learning. In the first part, we will study linear models and see examples of their extension to generalized linear models. Furthermore, we will discuss general principles of modelling. In the second part, we will study several modern non-parametric methods for regression and classification and evaluate their performance on different datasets. In addition, we will study modes of failure encountered when working with such flexible, non-parametric models. The emphasis throughout the module will be on principled, uncertainty-aware modelling.

Learning Outcome

- On successful completion of this module, students should be able to:
- Select an appropriate supervised learning method for a given application
- Use statistical models to extract information from processed data
- Design pipelines taking raw data as input and producing conclusions in context as output
- Interpret the output of statistical models in plain language
- Use suitable diagnostic tools to recognise common modes of failure of supervised learning algorithms

Module Content

- Data and models Linear modelling
- The modelling cycle Model flexibility: bias-variance,
- Regularisation
- Classification methods: logistic regression, LDA, QDA
- Resampling methods
- Model selection
- Decision and regression trees and random Forests
- Support Vector Machine

Year 1 – Term 3

MATH70102: Unsupervised Learning

Module Leader: Dr Mikko Pakkanen

Description

In this module we will introduce tools for performing different unsupervised learning tasks. The module will first focus on techniques for dimensionality reduction, including principal component analysis and its extensions, independent component analysis, and non-linear dimensionality reduction methods, such as t-SNE. Subsequently, we move on to parametric density estimation via maximum likelihood, including estimation of mixtures using the EM algorithm. In non-parametric density estimation, we study histograms and kernel density

estimators and learn how to set their tuning parameters. In the third part of the module, we study clustering methods, including K-means, K-medoids and hierarchical clustering, and learn how to evaluate the results of clustering quantitatively. In the final part, we learn about the distinction between anomaly and outlier detection and develop algorithms for these tasks.

Learning Outcomes

On successful completion of this module, students should be able to:

- Appreciate the differences between supervised and unsupervised learning
- Recognise the issues that arise from the curse of dimensionality
- Select, derive and apply algorithms for dimensionality reduction, density estimation and clustering
- Evaluate the performance of clustering techniques
- Deploy statistical approaches for anomaly and outlier detection

Module Content

In this module we will introduce tools for performing different unsupervised learning tasks. The lectures will focus on techniques for dimensionality reduction, parametric and non-parametric density estimation and clustering. Anomaly and outlier detection algorithms will also be discussed and developed.

MATH70100: Bayesian Methods and Computation

Module Leader: Prof. Nick Heard

Description

This specialisation introduces students to subjective probabilities and the Bayesian paradigm for making coherent individual decisions in the presence of uncertainty. The course will blend classical fundamental principles and mathematical rigour with a modern, high-level overview of a broad range of modern statistical techniques. Computer software packages will be introduced for implementing specific inferential procedures required for sophisticated Bayesian analyses.

Learning Outcome

On successful completion of this module, students should be able to:

- Distinguish the Bayesian interpretation of probability and decision making from other quantifications of uncertainty, variable estimation and prediction.
- Specify variable dependencies using graphical models.
- Follow the Bayesian inferential paradigm; calculate posterior probabilities, expectations and optimal decisions.
- Assess a problem and critically select an appropriate technique for estimation and inference for a Bayesian probability model. Distinguish parametric and nonparametric models and determine appropriateness of different model specifications.
- Use statistical software to apply Bayesian modelling techniques.

- Select an appropriate Bayesian model from a range of alternatives introduced throughout the course and follow the Bayesian paradigm for making inference.

Module Content

- Uncertainty and Decisions
- Prior and Likelihood Representation
- Graphical Modelling and Hierarchical Models
- Parametric Models
- Computational Inference
- Bayesian Software Packages
- Criticism and Model Choice
- Linear Models
- Nonparametric Models
- Nonparametric Regression

Year 2 – Term 1

MATH7104: Learning Agents

Module Leader: Dr. Kelly Zhang

Description

In an automated machine learning process, algorithms that make both inference and select decisions might be called learning agents. This module develops the expertise for taking machine learning beyond prediction process to formal decision-making processes. By contrasting issues that arise in the study of randomised controlled trials and formally designed experiments with issues related to the observational data, the module develops the theory and methodology of decision making through the theory of optimal decisions.

Learning Outcome

On successful completion of this module, students should be able to:

- Identify the conditions necessary to be able to use data to inform decision making
- Identify the pros and cons of using adaptive vs fixed experimental design and understand when it is appropriate to use each
- Reason about the exploration vs exploitation tradeoff and the algorithmic principles for sequential decision making under uncertainty
- Develop tools to address sequential decision making, and anticipate issues of deployment including the risks of training reinforcement learning algorithms online and how to mitigate them.

Module Content

- A/B testing
- Multi-armed bandits
- Reinforcement learning

MATH70103: Unstructured Data Analysis

Module Leader: Dr. Anthea Monod

Description

This module will provide the learner with the skills and knowledge to handle "unstructured' data, such as images, text and network data. Data science is replete with problems that involve unstructured data and this module develops methods for converting unstructured data to a more familiar "structured" form for use with standard Machine Learning methods as well as direct approaches with unstructured data. Examples will include natural language processing and network analysis.

Learning Outcome

On successful completion of this module, students should be able to:

- Manipulate unstructured data and covert to various mathematical representations
- Fluently uses a variety of methods for natural language programming, and understand their mathematical basis and shortcomings
- Manipulate network data, and use linear algebra approaches to conduct analysis
- Understand issues of model comparison and assessment unstructured data problems
- Sythensise data science pipelines for unstructured data analysis

Module Content

- Overview of modes of unstructured data
- Transformation of data for use with standard procedures
- Natural Language processing
- Network Analytics

Year 2 – Term 2

MATH70101 – Deep Learning

Module Leader: Dr. Kevin Webster

Description

This module teaches the building blocks of deep learning models, and how to design network architectures for specific applications, in both supervised and unsupervised contexts. It also covers practical skills in implementing neural networks. Students will learn how to design, implement, train and evaluate networks. A central focus of the module is on the mathematical and statistical foundations of some of the most sophisticated deep learning models, such as variational autoencoders (VAEs) and Bayesian methods for neural networks.

Learning Outcome

On successful completion of this module, students should be able to:

- Select appropriate deep learning model architectures for given supervised and unsupervised learning applications.

- Implement different neural network model architectures, loss functions and optimisers using the Keras framework (with either PyTorch or TensorFlow backend).
- Implement data and training pipelines for different types of neural networks using either the Keras framework (with either PyTorch or TensorFlow backend)
- Implement appropriate evaluation measures and model selection strategies for supervised and unsupervised applications

Module Content

- Deep learning fundamentals, layers, activation functions, loss functions
- Optimising deep learning models. Backpropagation algorithm
- Convolutional neural networks Sequence models. Recurrent neural networks
- VAEs, generative models
- Bayesian methods for deep learning

MATH70099: Big Data: Statistical scalability with PySpark

Module Leader: Dr. Francesco Sanna Passino

Description

The module consists of three components: statistical analysis at scale, distributed programming using MapReduce, Big Data analysis using PySpark. The first component covers theory on statistical scalability, and discusses topics such as sufficiency, Statistical Query Model, stochastic and distributed optimisation, stochastic variational inference Markov Chain Monte Carlo methods for tall data, and statistical analysis of streaming data. The second and third components cover practical aspects of handling Big Data, introducing two frameworks for analysis of large datasets: Hadoop and Spark.

Learning Outcome

On successful completion of this module, students should be able to:

- Extract, transform and load data using Hadoop Distributed File System in order to load data into and out of a big data environment.
- Use PySpark interface in order to interact with huge data sets.
- Perform EDA using PySpark in order to understand underpinning statistical properties of data set being analysed.
- Explain how underpinning statistical methodology can be applied to big data.
- Apply underpinning statistical methodology to big data using PySpark, in order to be able to produce statistical conclusion.
- Combine EDA methodology and PySpark knowledge to produce full statistical analysis on huge data sets.

Module Content

- Introduction to Big Data (history and characteristics) and statistical methods for Big Data analysis (divide and conquer, subsampling)
- Introduction to the Hadoop Distributed File System (HDFS) and MapReduce-based processing

- Statistical modelling in MapReduce (sufficiency and Statistical Query Model)
- Introduction to Resilient Distributed Datasets (RDDs) and DataFrames in PySpark for large scale Big Data Processing
- Statistical modelling with PySpark in the library ML
- User-defined functions and additional libraries in PySpark
- Optimisation with Big Data (stochastic gradient algorithms and stochastic variational inference)
- Markov Chain Monte Carlo and Big Data (divide-and-conquer MCMC and subsampling-based methods)
- Streaming data analysis (forgetting factor methods, exponentially weighted moving averages, change detection methods)

MATH70105 Research Project

Project Supervisor: Dr. Francesco Sanna Passino

Description

The module provides training in research on open-ended problems and gives the learner the opportunity to demonstrate the synthesis of the material taught over the programme. Research projects may be theoretical, methodological or applied depending on the interests of the learner. In all cases, the projects will require independent research, thinking and development, which will be scaffolded in a structured manner.

Learning Outcome

- Demonstrate the ability to research the background and details to new topics and analyse new data.
- Critically appraise new information.
- Design and conduct computational experiments to address new problems.
- Critically assess, and refine, the performance of methods for new problems.
- Report, by both presentation and document, the detailed findings of novel analysis, to both lay and expert technical audiences.