Multivariate Statistics and Machine Learning

Unit code:	MATH38161
Credit Rating:	10
Unit level:	Level 3
Teaching period(s):	Semester 1
Offered by	School of Mathematics
Available as a free choice unit?:	N

Requisites

Prerequisite

- MATH20701 - Probability 2 (Compulsory)
- MATH20802 - Statistical Methods (Compulsory)

Desirable

- Good working knowledge in the R statistical programming language

Aims

To familiarise students with the fundamental concepts and ideas underlying multivariate statistical data analysis methods and related supervised and unsupervised machine learning approaches for pattern recognition and classification, as well as with their practical implementation and application using the R statistical programming language.

Overview

Multivariate statistical models and methods are essential for analysing complex-structured and possibly high-dimensional data from any areas of science and industry, ranging from biology and medicine, and genetics to finance and sociology. Multivariate statistics also provides the foundation of many machine learning algorithms.

In the first part of this module covers the foundations of multivariate data analysis, e.g., multivariate random variables, covariance and correlation, and multivariate regression. In addition, related approaches such dimension reduction and latent variable models are discussed.

The second part of the course is concerned with multivariate approaches for statistical learning in supervised and unsupervised settings, including techniques from machine learning, and their application in pattern recognition, classification, and high-dimensional data analyis.

Learning outcomes

On successful completion of the course students will be able to:

- use the programming language R for multivariate data analysis and graphical presentation

- apply dimension reduction techniques such as PCA and CCA
- perform clustering and classification using tools from both statistics and machine learning
- make good choices among available parametric and nonparametric approaches
- analyse high-dimensional data sets with suitable regularisation techniques

Assessment methods

- Other - 50%
- Written exam - 50%

Assessment Further Information

- Coursework (2 written projects): weighting 50%
- End of semester examination: 1.5 hours weighting 50%

Syllabus

- Multivariate normal model (4 lectures): distributional properties, estimation of covariance and correlation matrix both in large and small sample settings (using likelihood and regularised/shrinkage estimation), connection with multivariate regression.
- Dimension reduction and latent variable models (4 lectures): whitening transformations, Principle Components Analysis (PCA), Canonical Correlation Analysis (CCA), Factor Analysis (FA)
- Unsupervised learning / clustering (4 lectures): model-based clustering (finite normal mixture models), algorithmic approaches (e.g. K-means, hierarchical clustering)
- Supervised learning / classification (6 lectures): Diagonal, Linear, and Quadratic Discriminant Analysis (DDA, LDA, QDA) and regularised versions for high-dimensional data analysis. Further approaches to classification (e.g. support vector machines).
- Nonlinear and Nonparametric models (4 lectures): decision trees, random forest

Recommended reading

- Berk, R. A 2016. *Statistical Learning from a Regression Perspective.* Second edition. Springer. Download PDF within UoM from https://link.springer.com/book/10.1007/978-3-319-44048-4
- Härdle, W.K., and L. Simar. 2015. *Applied Multivariate Statistical Analysis*. Fourth edition. Download within UoM from https://link.springer.com/book/10.1007/978-3-662-45171-7
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer. PDF freely available online from https://web.stanford.edu/~hastie/ElemStatLearn/

- James, G., D. Witten, T. Hastie and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer. PDF freely available online from http://www-bcf.usc.edu/~gareth/ISL/

Feedback methods

Computer labs will provide an opportunity for students to try out the methods on real data and to get feedback from the instructor. Courseworks and tutorials also provide an opportunity for students to receive feedback. Students can also get feedback on their understanding directly from the lecturer, for example during the lecturer's office hour or after class.

Study hours

- Lectures - 22 hours (11 x 2 hours)
- Tutorials - 15 hours (5 x 1 hour + 5 x 2 hours) - 5 tutorials will be computer based sessions
- Independent study hours - 63 hours

Timetable:

Week 1 – Multivariate normal model – no tutorial

Week 2 – Multivariate normal model – tutorial 1 (computer lab)

Week 3 – Dimension reduction – tutorial 2

Week 4 – Dimension reduction – tutorial 3 (computer lab)

Week 5 – Clustering – Tutorial 4

Week 6 – Reading week

Week 7 – Clustering – Tutorial 5

Week 8 – Classification – Tutorial 6 (computer lab)

Week 9 – Classification – Tutorial 7

Week 10 – Classification – Tutorial 8 (computer lab)

Week 11 – Nonlinear and nonparametric models – Tutorial 9

Week 12 – Nonlinear and nonparametric models – Tutorial 10 (computer lab)

Teaching staff

Korbinian Strimmer - Unit coordinator