Data Mining and Machine Learning II: Big Data and Unstructured Data

Course information sheet 2022-23

Full course, 10 weeks

This course introduces data mining and machine learning methods used in big data scenarios, most notably regularised regression, and also introduces methods for analysing networks and unstructured data.

Prerequisite Knowledge

Learners should have basic experience with the R programming language (e.g. data management and plotting).

Intended Learning Outcomes

By the end of this course learners will be able to:

- make appropriate use of informal and formal methods of network analysis;
- describe the challenges of the analysis of high-dimensional data and discuss, in a particular context, strategies for tackling big data problems;
- formulate and fit a regularised linear model, such as ridge regression, the LASSO and partial least-squares.
- infer statements about (conditional) independence from graphical models and factorisations of the joint distribution;
- describe methods for structural inference in graphical models and apply them in a given context;
- make appropriate use of informal and formal methods for quantitative text analysis.

Syllabus

Week 1 (sample material)

- Social network analysis
- Managing network data
- Visualising network data

Week 2

- Network summary statistics
- Node & edge level summaries
- Network models
- Stochastic block models

Week 3

- Stochastic block models
- Knowledge graphs
- Machine learning in graphs

Week 4

- Elastic nets
- Lasso and ridge regression
- Parameter estimation

Week 5

- Elastic nets
- Comparison between elastic nets and Lasso and ridge regression

Mid-term week break

Week 6

- Quantitative text analysis
- Preprocessing text data
- Simple summaries of texts

Week 7

- Peer assessment

Week 8

- Further quantitative text analysis
- Keyness, document similarity and keywords
- Sentiment analysis
- Topic models

Week 9

- Graphical models
- Graphics in R
- Undirected and directed graphs

Week 10

- Log-linear models
- Bayesian networks

“Really interesting and wide-ranging course which covers lots of interesting maching learning and data mining procedures.”

Online Learning

- Fortnightly live sessions with tutor(s)
- Weekly learning material (reading material, videos, exercises with model answers)
- Bookable one-to-one sessions with tutor(s)

Textbooks

Hastie, T & Tibshirani, R & Friedman, J (2009) Elements of statistical learning

Smola, A & Vishwanathan, S.V.N (2008) Introduction to machine learning

Assessment

(for credit only)

This will typically be made up of 4 pieces of assessment, including an online quiz, individual projects, and a peer assessment.

Software

To take our courses please use an up-to-date version of a standard browser (such as Google Chrome, Firefox, Safari, Internet Explorer or Microsoft Edge) and a PDF reader (such as Acrobat Reader). Learning material will be distributed through Moodle. We encourage all learners to install R and RStudio and we provide detailed installation instructions, but learners can also use free cloud-based services (RStudio Cloud). Learners need to install Zoom for participating in video conferencing sessions. We recommend the use of a head set for video conferencing sessions.

DATA ANALYTICS GLASGOW
School of Mathematics and Statistics
University of Glasgow
http://gla.ac.uk/mdatagov
http://gla.ai
Email:
maths-stats-analyticscpd@glasgow.ac.uk