ROBERT GORDON UNIVERSITY ABERDEEN

# CMM510 Data Mining

Pamela Johnston (SoCET)

p.johnston2@rgu.ac.uk

ROBERT GORDON UNIVERSITY ABERDEEN

**Statement for Audio and Video Learning Resources**

*Video and audio content at the University uses closed captions generated by automatic speech recognition (ASR). The ASR process is based on machine learning algorithms which automatically transcribe voice to text. According to our technology providers, this process is* ***approximately 70-90%*** *accurate depending on the quality of the audio, and consequently video and audio closed captions may include some transcription errors. It is therefore important to recognise that the original recording is the most accurate reflection of the content, and not the captions.*

*If you require accurate captions as part of your reasonable adjustments, please contact the Inclusion Centre to discuss your requirements.*

# Content

- Aims of CMM510 module
- Learning outcomes
- References
- Schedule
- Assessment

# Aim

- To provide students with an understanding of the main principles underlying **Data Mining** and **Machine Learning techniques** and the ability to **apply** current Data Mining and Machine Learning tools to real datasets.

# Learning Outcomes

On completion of this module, you should be able to:

- Critically discuss, compare and contrast the advantages and disadvantages of applying a specific data mining technique to a given learning task.
- Use industry standard tools to develop a data mining application tailored to a given learning task and evaluate the results obtained.
- Effectively interpret the results of learning through an understanding of the strengths and limitations of data mining technology and the selection of an appropriate evaluation technique.
- Demonstrate knowledge of the state-of-the-art in data mining.

# Critically discuss, compare and contrast the advantages and disadvantages of applying a specific data mining technique to a given learning task.

- Why did you pick that?
- Why not the other thing?
- This is **not** code!

# Use industry standard tools to develop a data mining application tailored to a given learning task and evaluate the results obtained.

- R
- R functions
- Did it work?

# Effectively interpret the results of learning through an understanding of the strengths and limitations of data mining technology and the selection of an appropriate evaluation technique

- How do you know it worked?
- Did it work well?
- What's wrong with it?
- *Maybe* R

# Demonstrate knowledge of the state-of-the-art in data mining.

- Name-check a few data mining techniques that give decent results and are in current use
- Know how they work under-the-hood
- Know something about what they replaced

# Topics

- Data mining concepts.
- Learning approaches: classification, regression (covered in CMM535), clustering and association rules.
- Bias-variance trade-off. Incorporating domain knowledge in learning.
- Advanced techniques for evaluating learned concepts. Calculation of confidence intervals for predictive performance. Comparison of data mining schemes.
- Boosting, bagging and stacking techniques.
- Applications.
- Legal, ethical, social and professional issues in data mining.

# References


- Bibliography
  - Sanjay Chakraborty, Sk Hafizul Islam, Debabrata Samantata, 2022, Data Classification and Incremental Clustering in Data Mining and Machine Learning, Springer.
  - Max Bramer, 2020. Principles of Data Mining. Springer.
  - Pavel Brazdil, Jan N. van Rijn, Carlos Soares, Joaquin Vanschoren, 2022, Metalearning : applications to automated machine learning and data mining (2nd Ed.). Springer.
  - David L. Olson, Georg Lauhoff, 2019, Descriptive Data Mining.
  - Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher Pal, 2017. Data Mining - Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufman.
  - David L Olson, 2019, Descriptive Data Mining. Springer.

- Software:
  - R, R: The R Project for Statistical Computing (r-project.org) https://www.r-project.org/ [accessed 04/09/2024]
  - Rstudio Posit | The Open-Source Data Science Company https://posit.co [accessed 04/09/2024]
  - W3 Schools R Tutorial (w3schools.com) https://www.w3schools.com/r [accessed 04/09/2024]

# Classes

- Lectures/tutorials:
  - Mondays, 12:00-13:00.
  - Note change next week!
- Lab sessions
  - Mondays 15:00-17:00.
  - Practical exercises attempted by students during session, supported by staff.
  - Solutions to labs published, no later than a week after the lab is released. Note that there are lots of possible solutions so if your solution is different to the published one and you are unsure about its correctness, ask.
  - A demonstration of lab work the week will be included if there is demand for it.

# Monday Holiday!

- 22nd September
- No classes. Classes rescheduled for later in the week.

# Sister module CMM535!

- Data Science Development
- Also in R
- First labs together just to learn R.

# Drop-in support sessions

- Support normal times (please check as these may change for individual weeks)
  - Mondays 2-3pmand Thursdays 10-11am on campus.
    - Let the School office (N447) know that you are there to see me.
    - Or send me a Teams message – if I'm at my desk, I'll get it.
    - I will come out to meet you.

# Assessment

- **Practical examination**
  - 3-hour assessment where you undertake a series of data mining tasks in a computer laboratory under exam conditions.
  - Access to **your own dropbox of notes**.
  - The **date** for the assessment is not yet known, but it will be in December.
  - Minimum grade D is required to pass the module.
  - There is one resit opportunity.
- A session on the practical examination will take place later in the year.
- The practical examination brief is on Moodle.