ISEL
INSTITUTO SUPERIOR
DE ENGENHARIA DE LISBOA

AMD

Aprendizagem e Mineração de Dados

(Machine Learning and Data Mining)

MEIM mandatory subject (1st semester)
MEIC elective subject (1st semester)
(winter-semester 2021/2022)

https://isel.pt/disciplinas/aprendizagem-e-mineracao-de-dados-meim

Synthesis

1. Overview
2. Goals and learning outcomes
3. Syllabus
4. Assessment
5. Planning
6. Teachers and contacts
7. Bibliography
8. Tools and resources

1. Overview

- AMD is founded on concepts from:
  - data bases and information systems
  - machine learning
  - programming
  - statistics

- AMD is preceeded by AA (Machine Learning) at LEIM

- AMD is optionally followed by MDLE (MEIM and MEIC)

1. Overview – “a step within a path”

		@LEIM	**@MEIM**	**@MEIC**
AA ML	Aprendizagem Automática Machine Learning	**M**		
AMD MLDM	**Aprendizagem e Mineração de Dados Machine Learning and Data Mining**		**M**	**O**
AAA AML	Aprendizagem Automática Avançada Advanced Machine Learning		**M**	
CDLE BDC	Computação de Dados em Larga Escala Big Data Computing		O	
MDLE BDM	Mineração de Dados em Larga Escala Big Data Mining		O	O

M: **M**andatory
O: **O**ptional

2. Goals and learning outcomes

AMD skills and competences to be developed by the students, are:

1. Build a "dataset" from different data-storage, e.g., relational model or text on the Web, considering its structure and semantics in order to draw hypotheses and interpret results.

2. Prepare data via de-normalization, assembling and discretization.

3. Explore the characteristics, options, benefits and limitations of supervised classification methods:

    a) with statistical support,

    b) based on the induction of decision trees,

    c) based on competitive learning.

2. Goals and learning outcomes (cont.)

AMD skills and competences to be developed by the students, are:

4. Introduce time series analysis; adapt dataset to apply (in this context) supervised classification.

5. Explore unsupervised methods based on instances.

6. Explore the methods that search for association rules and highlight the difference between those methods and the ones related to classification and clustering.

7. Evaluate learning via error estimation via training, validation and testing datasets; comparison of models and results presentation.

3. Syllabus

I. Generate and export "dataset" from relational model and Web data; numerical and nominal domains and missing values.

II. Unsupervised and supervised approaches to discretization.

III. Classification with Bayes and Laplace estimators.

IV. Induction of decision trees; intrinsic information, information gain, gain ratio and Gini index; nominal attributes; methods ID3 and C4.5; overfitting and (pre/post)-tree pruning; learning-vector-quantization, atraction and repulsion and learning rate.

3. Syllabus (cont.)

V. Clustering and classification based on instances; distance functions with numeric and nominal domain and missing values; neighborhood searching with KD-Tree and support to kNN (classification) and K-means (clustering).

VI. Association rules; market-basket analysis, rule-space and assessment (support and confidence); APRIORI and H-Mine.

VII. Error rate and training, validation and testing sets; cross-validation and bootstrap; errors and costs; confusion matrix, Kappa and ROC (single/multi-class).

4. Assessment – calculation method

- **Final-grade** =
  0,5*theoretical-grade + 0,5*practical-grade

- **Theoretical-grade (>= 9,5 grade)**
  - individual via written final-exam
  - regular or appeal evaluation moments

- **Practical-grade (>= 9,5 grade)**
  - worksheets resolutions during the semester class's time period
    - exercises to explore and consolidate the comprehension of theoretical concepts,
    - exercises that integrate into the final project,
  - final project (to start during the semester class's time period)
  - discussion of final project and practical classes' worksheets

4. Assessment – important remarks

- **Each submitted work will be previously analyzed using a computer plagiarism detection tool.**

- **The existence of plagiarism in practical work will lead to:**
  - **nullification of all the involved work,**
  - **immediate failure, in this subject, of all the involved students.**

- Only works whose authors coincide with the constitution of groups in the Moodle system will be accepted.

- Any withdrawals must be communicated to the class teacher.

- Delivery of practical work will not be accepted after the deadline.

5. Planning

week	**activity**
4/oct	begin of the schoolar term
18/oct – 22/oct	kick-off of first final-project (final-project-A)
22/nov – 26/nov	kick-off of second final-project (final-project-A1)
06/dec – 10/dec	kick-off of third final-project (final-project-B)
22/jan	end of the scholar term
jan/feb	completion of the discussion and the final-exam

6. Teachers and contacts

(day-time) teacher MM1D-MI1D - Artur Ferreira

- aferreira@deetc.isel.ipl.pt artur.ferreira@isel.pt

(night-time) teacher MM1N-MI1N - Paulo Trigo (UC responsible)

- ptrigo@deetc.isel.ipl.pt paulo.trigo@isel.pt

UC Moodle pages

AMD-UCC (common contents)

https://2122moodle.isel.pt/course/view.php?id=4462

AMD-MM1D-MI1D (day-time class)

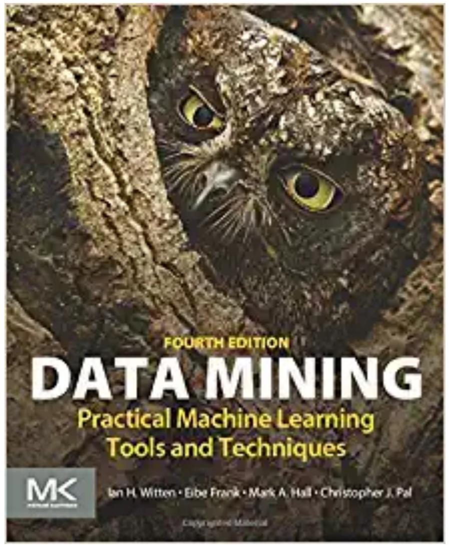
https://2122moodle.isel.pt/course/view.php?id=4460

AMD-MM1N-MI1N (night-time class)

https://2122moodle.isel.pt/course/view.php?id=4461

7. Bibliography

Witten, H. I., Frank, E., Hall, M. A., and Pal, C. J. (2016). Data Mining – Practical Machine Learning Tools and Techniques; (4th ed.);
Morgan-Kaufmann


FOURTH EDITION
DATA MINING
Practical Machine Learning Tools and Techniques
MK
Ian H. Witten • Eibe Frank • Mark A. Hall • Christopher J. Pal

Orange Data Mining Library Documentation. (2018); Orange Data Mining.

https://readthedocs.org/projects/orange-data-mining-library/downloads/pdf/latest/

8. Tools and resources

- Python (programming language)
- Orange DM (data mining graphical and programmatic tool)
- SQL (Structured Query Language)
- PostgreSQL (database management system)