Ficha de Unidade Curricular A3ES
Aprendizagem e Mineração de Dados
Mestrado em Engenharia Informática e de Computadores
2025-26

1. Designação da unidade curricular

[2773] Aprendizagem e Mineração de Dados / Machine Learning and Data Mining

2. Sigla da área científica em que se insere	IC, n/d
3. Duração	Unidade Curricular Semestral
4. Horas de trabalho	162h 00m
5. Horas de contacto	Total: 60h 00m, h 00m das quais T: 22h 30m, h 00m | TP: 7h 30m, h 00m | P: 30h 00m, h 00m
6. % Horas de contacto a distância	Sem horas de contacto à distância
7. ECTS	6
8. Docente responsável e respetiva carga letiva na Unidade Curricular	[1457] Paulo Manuel Trigo Cândido da Silva | Horas Previstas: 60 horas
9. Outros docentes e respetivas cargas letivas na unidade curricular	[1375] Artur Jorge Ferreira | Horas Previstas: 60 horas

10. Objetivos de aprendizagem e a sua compatibilidade com o método de ensino (conhecimentos, aptidões e competências a desenvolver pelos estudantes).

1. Construir dataset a partir de repositórios, e.g., modelo relacional ou texto Web, considerando a estrutura e semântica, com o objetivo de colocar hipóteses e interpretar resultados

2. Preparar dados via des-normalização, composição e discretização

3. Explorar as características, opções, vantagens e limitações dos métodos de classificação: a) de suporte estatístico, b) baseados na indução de árvores de decisão, c) baseados em aprendizagem competitiva

4. Introduzir a análise de séries temporais; adaptação de datastet para aplicar (neste contexto) métodos de classificação supervisionada

5. Explorar métodos não-supervisionados baseados em instâncias

6. Explorar os métodos de procura de regras de associação e evidenciar a diferença em relação à classificação e agrupamento

7. Avaliar a aprendizagem via estimação de erro suportado nas noções de conjuntos de treino, validação e teste; comparação de modelos e apresentação de resultados.

10. Intended Learning objectives and their compatibility with the teaching method (knowledge, skills and competences by the students).

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

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

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 supported on the concepts of training, validation and testing sets; comparison of models and results presentation.

11. Conteúdos programáticos

I. Gerar e exportar dataset a partir do modelo relacional e dados Web; domínios numéricos e nominais e valores omissos.

II. Abordagens não-supervisionadas e supervisionadas à discretização.

III. Classificação com formulação de Bayes e estimadores Laplace.

IV. Indução de árvores de decisão; informação intrínseca, ganho de informação, rácio do ganho e gini index; métodos ID3 e C4.5; sobre-ajuste e poda (pre/post-pruning); learning-vector-quantization, operadores de atração e repulsa e passo-de-aprendizagem

V. Agrupamento e classificação baseada em instâncias; funções distância com atributos numérico, nominal e omisso; procura de vizinhos com KD-Tree e suporte ao kNN (classificação) e K-means (clustering).

VI. Regras de associação; market-basket analysis, rule-space e avaliação (suporte e confiança); método APRIORI e H-Mine.

VII. Taxa de erro e conjuntos de treino, validação e teste; validação cruzada e bootstrap; erros e custos; matriz confusão, Kappa e ROC (uni/multi-classe).

11. 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 operators and learning rate

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).

12. Demonstração da coerência dos conteúdos programáticos com os objetivos de aprendizagem da unidade curricular

Esta UC percorre as fases do processo de data-mining seguindo o essencial proposto pelo Cross-Industry Standard Process for Data Mining (CRISP-DM), no entanto a ênfase está menos no processo e mais no aprofundar e aplicar as técnicas procurando alinhar os diferentes tipos de problema (classificação, agrupamento, regras associação) com cada grupo de métodos (baseada em estatística, indução de árvores ou instâncias) considerando o impacto da escolha dos atributos (features), do domínio dos atributos e dos (eventuais) valores omissos.

A abordagem reflete-se na distribuição dos conteúdos programáticos. Alinhando com o CRISP-DM temos: (a) itens I a II dedicados ao business understanding e data preparation, (b) itens III, IV, V e VI fase de modeling com ênfase na caraterização fina de alguns algoritmos, (c) item VI fase de evaluation incluindo o reporte e apresentação de resultados e conclusões, e (d) uma fase de deployment concretizada em contexto de projeto final da UC.

12. Evidence of the syllabus coherence with the curricular unit´s intended learning outcomes

This UC covers the various stages data-mining process following the essentials proposed by Cross-Industry Standard Process for Data Mining (CRISP-DM), however, emphasis is less on process and more on deepening and implementing techniques aiming for aligning the problem types (classification, clustering, association rules) with each group of methods (statistics based, tree induction or instances) always considering impact of attributes' choice (features), domain of attributes and (possible) existence of missing values.

This approach is mirrored in the distribution of the syllabus. Lining up with CRISP-DM we have: (a) items I to II aimed at business understanding and data preparation, (b) items III, IV, V and VI dedicated to modeling with emphasis on the characterization of algorithms, (c) Item VI dedicated to evaluation including reporting and presentation of results and conclusions, and (d) the deployment stage implemented in the context of the final project of the UC.

13. Metodologias de ensino e de aprendizagem específicas da unidade curricular articuladas com o modelo pedagógico

No ponto 5 (acima), T corresponde à exposição de conceitos e exploração de bases suportados no estudo de casos; TP corresponde à realização de exercícios práticos guiados por etapas bem-definidas; PL corresponde à realização de trabalho prático a partir de um enunciado que estabelece os pressupostos e alinha os passos para alcançar os objetivos, tentando ainda promover a autonomia (do estudante) e a capacidade para analisar e concluir com base em resultados gerados via experimentação. Na TP e na PL (TP\PL) a perspectiva prática (P) concretiza-se, em geral, com recurso ao computador.

T: 22,5h (1.5h*15s). Apresentação e discussão de conceitos teóricos com recurso a exemplos práticos. Caracterização e análise dos casos de aplicação a desenvolver na (próxima) TP\PL.

TP\PL: 37.5h ([0.5h\aulaTP + 2h\aulaPL]*15s). Em cada aula há uma ficha de problemas sobre o tema da (anterior) aula T. Há dois tipos de fichas: a) exercícios TP para explorar e consolidar a compreensão de conceitos teóricos, e b) exercícios PL cuja resolução contribui com uma componente a integrar no projeto final.

Realização autónoma de projeto final (102h) com suporte das aulas TP\PL e do docente.

13. Teaching and learning methodologies specific to the curricular unit articulated with the pedagogical model

In item 5 (above), T corresponds to the exposition of concepts and exploration of basis supported on case-studies; TP corresponds completion of practical exercises guided by well-defined stages; PL corresponds to the accomplishment of practical work from a statement that establishes the assumptions and aligns the steps to reach the goals, also trying to promote the autonomy (of the student) and the ability to analyze and conclude based on results generated throughout experimentation. In TP and PL (TP\PL) the practical perspective (P) is usually implemented using the computer.

T: 22,5h (1.5h*15s). Presentation and discussion of concepts via practical examples. Characterization and analysis of practical cases to be developed in the (next) TP\PL lecture.

TP\PL: 37.5h ([0.5h\aulaTP + 2h\aulaPL]*15s). Each lecture presents a problems? worksheet related with (previous) T lecture subjects. There are two types of worksheets: a) TP exercises to explore and consolidate the comprehension of theoretical concepts, and b) PL exercises that integrate into the final project.

Autonomous realization of the final project (102h) with support from TP\PL classes and the teacher.

14. Avaliação

A avaliação é distribuída com exame final. Todos os elementos de avaliação são pedagogicamente fundamentais (cada elemento da avaliação distribuída tem nota mínima de 9.5).

T: individual via exame escrito

P: individual via discussão com grupo (projeto x 0.35 + relatório x 0.3 + discussão x 0.35)

Nota Final (NF): (T + P) / 2

Aprovação: T >= 9.5 e P >= 9.5 e NF >= 9.5.

14. Assessment

Assessment is distributed with a final exam. All assessment elements are pedagogically fundamental (each component of the distributed assessment has a minimum mark of 9.5).

T: individual via written exam

P: individual via group discussion (project x 0.35 + report x 0.3 + discussion x 0.35)

Final Grade (NF): (T + P) / 2

Approval: T >= 9.5 and P >= 9.5 and NF >= 9.5.

15. Demonstração da coerência das metodologias de ensino com os objetivos de aprendizagem da unidade curricular

Nas aulas T são expostos os conteúdos programáticos, focados em alcançar os objectivos de aprendizagem 1 a 7, acompanhados com problemas ilustrativos a ser resolvidos pelos alunos, na sala, antes de apresentada a solução. No fim de cada aula T é indicado o objectivo da próxima aula TP\PL e é colocado no moodle a respectiva ficha de problemas.

Nas aulas TP\PL abordam-se os vários objectivos e as competências para desenhar uma solução envolvendo caracterização do problema, construção de dataset, preparação dos dados, aplicação de métodos e avaliação no sentido de justificar o modelo a fazer deploy.

A perspetiva TP\PL é a de ir amadurecendo, de modo incremental, a compreensão das técnicas num contexto de experimentação e visando alcançar determinada funcionalidade especifica. A experimentação recorre a ambientes de código fonte aberto - Orange DataMining, PostgreSQL, e linguagens de programação Python, SQL.

15. Evidence of the teaching methodologies coherence with the curricular unit´s intended learning outcomes

In T classes syllabus content is presented focused in achieving the learning outcomes 1 to 7, along with illustrative problems to be solved by the students, in the classroom, prior to the solution presentation. At the end of each T class it is described the goal of the next TP\PL class and the corresponding worksheet is made available in the moodle.

In TP\PL classes all the learning outcomes are explored and the skills to design a solution involving the characterization of the problem, dataset construction, data preparation, implementation and evaluation methods in order to justify the model to deploy.

The perspective TP\PL is to incrementally get a matured understanding of the techniques in an experimental context and aiming to achieve a specific functionality. The experimentation resorts to open source environments - Orange DataMining, PostgreSQL, and programming languages, Python, SQL.

16. Bibliografia de consulta/existência obrigatória

1. 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.

2. Orange Data Mining Library Documentation. (2023). Orange Data Mining.

3. Brownlee, J. (2017). Machine Learning Mastery with Python. eBook.

4. Hastie, T., Tibshirani, R., Friedma, J. (2017). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

5. Goodfellow, I., Bengio Y., and Courville, A. (2016). Deep Learning. MIT Press. www.deeplearningbook.org

6. Cord, M., and Cunningham, P. (2008). Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Springer.

17. Observações

Unidade Curricular Obrigatória, Unidade Curricular Opcional
Unidade Curricular comum ao(s) curso(s) de MEIM

Data de aprovação em CTC: Data de aprovação em CTC:

Data de aprovação em CP: Data de aprovação em CP: