

Ficha de Unidade Curricular – (Versão A3ES 2018-2023)

1. Caracterização da Unidade Curricular.

1.1. Designação da unidade curricular (1.000 carateres).

Aprendizagem e Mineração de Dados / Machine Learning and Data Mining

1.2. Sigla da área científica em que se insere (100 carateres).

IC

1.3. Duração¹ (100 carateres).

Semestral

1.4. Horas de trabalho² (100 carateres).

162h (6ECTS * 27h/ECTS)

1.5. Horas de contacto³ (100 carateres).

67,5; T: 22,5h; TP: 15h; PL: 30

1.6. ECTS (100 carateres).

6

1.7. Observações⁴ (1.000 carateres).

No ponto 1.5 (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 de 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.

UC comum com outros cursos

UC é obrigatória em MERCM

UC é optativa em MEIC

1.7. Remarks (1.000 carateres).

In item 1.5 (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.

UC is mutual with other courses

UC is mandatory in MERCM

UC is optional in MEIC

2. Docente responsável e respetiva carga letiva na Unidade Curricular (*preencher o nome completo*) (1.000 carateres).

Paulo Manuel Trigo Cândido da Silva; 67,5 horas de contacto

3. Outros docentes e respetivas cargas letivas na unidade curricular (1.000 carateres).

Não se aplica / Not applicable

4. Objetivos de aprendizagem (conhecimentos, aptidões e competências a desenvolver pelos estudantes). (1.000 carateres).

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 desnormalização, 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 dataset 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.

4. Intended learning outcomes (knowledge, skills and competences to be developed by the students). (1.000 characters).

1. Build a "dataset" from different datastorage, 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 denormalization, 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 dataset 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.

5. Conteúdos programáticos (1.000 caracteres).

- 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; sobreajuste e poda ("pre/postpruning"); "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 omissos; procura de vizinhos com KDTree e suporte ao kNN (classificação) e K-means ("clustering").
- VI. Regras de associação; "marketbasket analysis", "rulespace" 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/multiclasse).

5. Syllabus (1.000 characters).

- 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, attraction 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 KDTree and support to kNN (classification) and Kmeans (clustering).
- VI. Association rules; marketbasket analysis, rulespace and assessment (support and confidence); APRIORI and HMine.
- VII. Error rate and training, validation and testing sets; crossvalidation and bootstrap; errors and costs; confusion matrix, Kappa and ROC (single/multiclass).

6. Demonstração da coerência dos conteúdos programáticos com os objetivos de aprendizagem da unidade curricular (1.000 caracteres).

Esta UC percorre as fases do processo de “datamining” seguindo o essencial proposto pelo “CrossIndustry Standard Process for Data Mining” (CRISPDM), 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 refletese na distribuição dos conteúdos programáticos. Alinhando com o CRISPDM 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 caracterização 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.

6. Evidence of the syllabus coherence with the curricular unit’s intended learning outcomes (1.000 characters).

This UC covers the various stages "datamining" process following the essentials proposed by "CrossIndustry Standard Process for Data Mining" (CRISPDM), 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 CRISPDM 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.

7. Metodologias de ensino (avaliação incluída) (1.000 carateres).

T: 22,5h (15aula*1,5h/aula). 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: 45h (15aula*(1h/aulaTP + 2h/aulaPL)). Cada aula uma ficha com 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. A resolução de cada ficha é enviada ao docente (até aula seguinte); dificuldades, de cada ficha, são analisadas em cada aula e o conjunto das fichas é avaliado na discussão do projeto final.

Realização autónoma de projeto final (94,5h) com suporte da aulas TP\PL e do docente.

Trabalhos em grupo (preferência 2 alunos).

Avaliação:

T: individual via exame escrito

P: individual via discussão com grupo (fichas e projeto final)

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

Aprovação: $T \geq 9,5$ e $P \geq 9,5$ e $NF \geq 9,5$.

7. Teaching methodologies (including assessment) (1.000 characters).

T: 22,5h (15class*1,5h/class). 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: 45h (15class*(1h/classTP + 2h/classPL)). Each lecture presents a problems' worksheet related with (previous) T lecture subjects. Two types of worksheets: a) exercises TP to explore and consolidate the comprehension of theoretical concepts, and b) exercises PL that integrate into the final project. Worksheet resolution sent to teacher (until next class); difficulties, in worksheet, are analyzed in each class and the set of all worksheets is assessed in final project discussion.

Autonomous final project realization (94,5h) with the support of TP\PL classes and teacher.

Work done in teams (likely 2 students).

Assessment:

T: individual via written exam

P: individual via vivavoce exam with group (worksheets and final project)

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

Approval: T >= 9,5 and P >= 9,5 and NF >= 9,5.

8. Demonstração da coerência das metodologias de ensino com os objetivos de aprendizagem da unidade curricular (3.000 caracteres).

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

8. Evidence of the teaching methodologies coherence with the curricular unit’s intended learning outcomes (3.000 characters).

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.

9. Bibliografia de consulta/existência obrigatória (1.000 caracteres).

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 [pdf]. (2018). 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.

¹ Anual, semestral, trimestral, ...

² Número total de horas de trabalho.

³ Discriminadas por tipo de metodologia adotado (T - Ensino teórico; TP - Ensino teórico-prático; PL - Ensino prático e laboratorial; TC - Trabalho de campo; S - Seminário; E - Estágio; OT - Orientação tutorial; O - Outro).

⁴ Assinalar sempre que a unidade curricular seja optativa.