Ficha de Unidade Curricular

1. Caracterização da Unidade Curricular.

1.1. Designação da unidade curricular ***(1.000 carateres).***
Elementos de Aprendizagem Estatística / Statistical Learning Elements

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

1.3. Duração[1] ***(100 carateres).***
Semestral

1.4. Horas de trabalho[2] ***(100 carateres).***
162

1.5. Horas de contacto[3] ***(100 carateres).***
TP: 45; PL:22.5

1.6. ECTS ***(100 carateres).***
6

1.7. Observações[4] ***(1.000 carateres).***

1.7. Remarks ***(1.000 carateres).***

2. Docente responsável e respetiva carga letiva na Unidade Curricular ***(preencher o nome completo) (1.000 carateres).***
Sandra Maria da Silva Figueiredo Aleixo, 22.5h

3. Outros docentes e respetivas cargas letivas na unidade curricular ***(1.000 carateres).***
Carlos José Brás Geraldes, 22.5 h
Iola Maria Silvério Pinto, 22.5h

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

Esta UC introduz os algoritmos e métodos estatísticos para mineração de dados. A sua natureza interdisciplinar combina tópicos de estatística, bases de dados e ciência da computação. Os objetivos são:

1. Identificar as fases de um projeto de *Data Science*. Conhecer os conceitos e técnicas estatísticas de aprendizagem automática (AA), exemplificando as suas aplicações e funcionalidades
2. Identificar os vários tipos de dados de input e conhecer os métodos para a sua preparação e pré-processamento, bem como as principais técnicas para transformação e redução da dimensionalidade dos dados
3. Conhecer os fundamentos teóricos matemáticos dos métodos de AA apresentados, saber manuseá-los, bem como identificar e interpretar várias formas e tipos de resultados
4. Avaliar os resultados obtidos com as técnicas de AA, usando e interpretando as medidas de desempenho
5. Usar softwares apropriados para resolver vários desafios
6. Realizar na integra um projeto de *Data Science* usando metodologias adequadas

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

This CU introduces statistical data mining algorithms and methodologies. It´s interdisciplinary nature combines topics of statistics, databases, and computer science. The intended outcomes are:

1. To describe the several stages of a Data Science project. To know the concepts and statistical techniques of machine learning (ML), exemplifying its applications and functionalities
2. To identify the several types of input data and to know the methodologies for its preparation and preprocessing, as well as the main techniques for transformation and reduction of data dimensionality
3 To know the mathematical theoretical foundations associated to the ML methods, to know how to handle them, as well as to identify and interpret the several forms and types of results

4. To evaluate the results obtained with the ML techniques, using, and interpreting the performance measures
5. Use appropriate software for solving several challenges
6. To complete a Data Science project using appropriate methodologies

5. Conteúdos programáticos ***(1.000 carateres).***

1. Introdução. Fases de um projeto de Data Science. Conceitos e ferramentas de mineração de dados e aprendizagem automática. Input: tipos de dados e preparação de dados. Output: tipos de representação
2. Classificação. Método do Vizinho Mais Próximo (KNN)
3. Classificadores de Bayes. Classificador *Naive Bayes*
4. Estimação da Densidade
5. Desempenho do Classificador
6. Árvores de Decisão para a Classificação
7. Classificador de Máquinas de Vetores de Suporte
8. Análise de Componentes Principais
9. Análise de Clusters
10. Análise de Componentes Independentes
11. Classificação com Análise Discriminante
12. Avaliação de Desempenho dos Classificadores

5. Syllabus ***(1.000 characters).***

1. Introduction. Phases of a Data Science project. Data mining and machine learning concepts and tools. Input: data types and data preparation. Output: representation types
2. Classification. Nearest Neighbor Method (KNN)
3. Bayes classifiers. Naive Bayes Classifier
4. Density estimation methods
5. Classifier Performance
6. Decision Trees for Classification
7. Support Vector Machine Classifier
8. Principal Component Analysis
9. Cluster Analysis
10. Independent Component Analysis
11. Classification with Discriminant Analysis
12. Classifier Performance Assessment

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

Os pontos 1 e 2 dos conteúdos programáticos pretende alcançar o ponto 1 dos objetivos
Os pontos 2, 3, 4, 6, 7, 9, 10 e 11 dos conteúdos programáticos apresentam os conceitos necessários para atingir o ponto 3 dos objetivos
Os pontos 5 e 12 dos conteúdos programáticos pretendem alcançar o ponto 4 dos objetivos
A totalidade dos pontos dos conteúdos programáticos permitem atingir e consolidar os objetivos 5 e 6

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

Points 1 and 2 of the syllabus aim to achieve point 1 of the objectives
Points 2, 3, 4, 6, 7, 9, 10 and 11 of the syllabus present the concepts necessary to achieve point 3 of the objectives
Points 5 and 12 of the syllabus aim to achieve point 4 of the objectives
All the points of the syllabus allow achieving and consolidating objectives 5 and 6

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

As aulas são teórico-práticas. A parte teórica baseia-se no formalismo matemático e na compreensão dos algoritmos subjacentes à implementação dos métodos de AA, promovendo o desenvolvimento de capacidade crítica no processo de modelação. Os alunos serão familiarizados com as técnicas de aprendizagem estatística, que incluem a inferência e a interpretação dos modelos de classificação. Na parte prática são implementados computacionalmente os algoritmos de AA e os procedimentos abordados na parte teórica, usando *software* livre (preferencialmente o R), com base em casos que podem ser reais ou simulados. São disponibilizados elementos de apoio aos conteúdos programáticos. A avaliação tem 2 componentes: a teórica é constituída por um exame (nota mínima de 9,5 valores) e a prática compreende um projeto de grupo com apresentação e discussão obrigatória (nota mínima de 9,5 valores). A nota final, NF, é dada por: NF=0,5 NT+0,5 NP, onde NT é a nota da parte teórica e NP a nota da parte prática.

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

Classes are theoretical-practical. The theoretical part is based on the mathematical formalism, in understanding the algorithms underlying the implementation of ML methods, allowing the development of critical capacity in the modeling process. Students will be familiarized with the techniques of statistical learning, including inference and classification model interpretation. In the practical part, the ML models and procedures covered in the theoretical part are computationally implemented, using free software (preferably R) based on cases that can be real or simulated. Elements of support for the syllabus are made available. Knowledge assessment comprises 2 components: the theoretical consists of an exam (minimum score of 9.5) and the practical comprises a group's project with mandatory presentation and discussion (minimum score of 9.5). The final grade, NF, is given by: NF = 0.5 NT + 0.5 NP, where NT is the grade of the theoretical part and NP the grade of the practical part.

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

A aprendizagem estatística é focada fundamentalmente na formalização de todos os aspetos do processo de mineração de dados, principalmente naqueles que permitem a interpretação e avaliação dos resultados obtidos pelo modelo de classificação. É importante que o aluno adquira competências na utilização do raciocínio estatístico para resolver problemas de modelação de complexidade mais elevada.

As metodologias de ensino são consistentes com os objetivos da unidade curricular, uma vez que na parte teórica é ensinado o formalismo estatístico necessário para o desenvolvimento da capacidade crítica na escolha e construção de um modelo de classificação. Adicionalmente, na parte prática, o aluno pode aplicar as aptidões já mencionadas para resolver problemas reais ou simulados (próximos da realidade).

A implementação computacional dos processos envolvidos, com recurso à utilização de um *software* livre, irá possibilitar ao aluno o desenvolvimento das suas próprias ferramentas bem como um melhor entendimento da resolução prática de problemas.

A avaliação da aprendizagem com base num exame permitirá aferir os conhecimentos e competências individuais adquiridas pelo aluno. O projeto de grupo permitirá avaliar a capacidade cooperativa na resolução dos problemas.

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

Statistical Learning is very focused on the formalization of all aspects of the data mining process, especially those that allow the interpretation and performance evaluation. It is important that the student acquires skills in the use of statistical reasoning to solve modelling problems of higher complexity.

The teaching methodologies are consistent with the objectives of the curricular unit, given that in the theoretical part the statistical formalism necessary for the critical capacity for the choice and development of a classification

model is taught and in the practical part, the student can apply the skills mentioned above for solving real or simulated examples (close to reality).

The teaching methodologies are consistent with the objectives of the curricular unit, since in the theoretical part is taught the statistical formalism needed for the development of critical capacity in the choice and construction of a model. Additionally, in the practical part, the student can apply the skills already mentioned to solve real or simulated examples (close to reality). The emphasis on issues related to interpretability and inference will result in a holistic understanding of the problem to be solved, which will be very useful in real situations of the student's future professional life.

The computational implementation of the processes involved, using free software, will enable students to develop their own tools as well as a better understanding of practical problem solving.

The assessment of learning based on an exam will allow the evaluation of individual knowledge and skills acquired by the student. The group's project will allow assessing the cooperative capacity in solving problems.

9. Bibliografia de consulta/existência obrigatória ***(1.000 carateres).***

1. Bishop, C. M., Pattern Recognition and Machine Learning, Springer (2006).
2. Duda, R. O., Hart, P. E. and Stork, D. G., Pattern Classification, Wiley (2001).
3. Hastie, T., Tibshirani, R. and Friedman, J., The elements of statistical learning. Springer (2017).
4. Lantz, B., Machine Learning with R, Packt (2013).
5. Muller, A. and Guido, S., Introduction to Machine Learning with Python, O'Reilly (2017)
6. Murphy, K., Machine Learning: A Probabilistic Perspective, MIT Press (2012).
7. Witten, I. H., Frank, E. and Hall M. A., Data mining: practical machine learning tools and techniques. Morgan Kaufmann (2011).

[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.