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

1. Caracterização da Unidade Curricular.

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

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 OT: 5

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

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

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

2. Docente responsável e respetiva carga letiva na Unidade Curricular ***(preencher o nome completo) (1.000 carateres).***
Carlos José Brás Geraldes, 67.5 h.

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

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

1. Aplicar técnicas de modelação sob o quadro geral da teoria da aprendizagem estatística
2. Formalizar modelos orientados para a predição e inferência
3. Aplicar os conceitos de predição e inferência para resolução de problemas em contextos específicos
4. Selecionar e avaliar os modelos com base nos respetivos desempenhos
5. Interpretar os resultados bem como as associações entre as variáveis a partir dos modelos aplicados no contexto do problema em estudo
6. Implementar computacionalmente as técnicas de modelação e interpretação utilizando o *software* mais apropriado

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

1. To apply modeling techniques under the general framework of statistical learning theory
2. To formalize models oriented to prediction and inference
3. To apply the concepts of prediction and inference to solve problems in specific contexts
4. To select and evaluate the models based on their performance
5. To interpret the results as well as the associations between variables from the models in the problem context under study
6. To implement, computationally, the modeling and interpretation techniques using appropriate software

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

1. Aspetos da aprendizagem supervisionada.
    1.1. Introdução à aprendizagem estatística. Aprendizagem supervisionada. Definição de variáveis e adaptação de modelos.
    1.2. Classificadores generativos (distribuição de probabilidade conjunta). Classificadores discriminativos (distribuição de probabilidade condicional). Aproximação de funções.
2. Métodos lineares para regressão e classificação.
    2.1. Revisão dos conceitos de regressão linear. Estimação de parâmetros através do erro quadrático mínimo e da máxima verosimilhança. Teorema de Gauss-Markov.
    2.2. Seleção de variáveis. Métodos *forward and backward-stepwise*. Regressões *Ridge* e Lasso (*Least absolute shrinkage and selection operator*).
    2.3. Regressão logística.
3. Métodos não lineares para regressão e classificação.
    3.1. Suavização. Suavizadores de médias móveis e retas móveis. Suavizadores *Kernel*. *Splines* suavizadores. Regressão *Kernel*. Regressão polinomial local. Regressão por *Splines.*
    3.2. Modelo aditivo generalizado (GAM). Estimação das funções parciais. Estimação da função de razão de possibilidades.
    3.3. Árvores de classificação e regressão.
    3.4. Máquinas de vetores de suporte.
    3.5. Redes neuronais. O perceptrão multicamada. Redes neuronais aditivas generalizadas.
4. Avaliação e seleção de modelos.
    4.1. Conceito de generalização. Complexidade do modelo. Compromisso viés-variância.
    4.2. Teoria de Vapnik-Chervonenkis. Métodos para reduzir o otimismo do erro de treino - validação cruzada e *bootstrap.*
5. Inferência e interpretabilidade dos modelos.
    5.1. Métodos bootstrap e máxima verosimilhança.
    5.2. Algoritmo EM.
    5.3. Interpretação dos modelos de regressão e classificação lineares.
    5.4. Interpretação das funções parciais de um GAM. Interpretação da função da razão de possibilidades.
    5.5. Método para a interpretação local e agnóstica de modelos – LIME (Local Interpretable Model-Agnostic Explanations).
6. Aprendizagem não supervisionada.
    6.1. Revisão sobre os conceitos de análise de agrupamentos (*Clusters*) e de análise em componentes principais (*Principal Component Analysis* - PCA).
    6.2. Análise em componentes independentes.
    6.3. Mapas Auto-Organizados (*Self Organizing Maps* - SOM).

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

1. Aspects of supervised learning.
    1.1. Introduction to statistical learning. Supervised learning. Variable definition and model adaptation.
    1.2. Generative classifiers (joint probability distribution). Discriminative classifiers (conditional probability distribution). Function approximation.

2. Linear methods for regression and classification.
    2.1. Review of linear regression concepts. Parameter estimation using the minimum quadratic error and maximum likelihood. Gauss-Markov theorem.
    2.2. Variable selection; *Forward and Backward-Stepwise*; *Ridge* and Lasso (Least absolute shrinkage and selection operator) regressions.
    2.3. Logistic regression.
3. Nonlinear methods for regression and classification.
    3.1. Smoothing. Moving average. Kernel smoothers. Smoothing splines. Kernel regression. Local polynomial regression. Regression Splines.
    3.2. Generalized additive model (GAM). Partial function estimation; Estimation of the odds ratio function.
    3.3. Regression and classification trees.
    3.4. Support vector machine.
    3.5. Neural Networks. The multi-layer perceptron. Generalized additive neural networks.
4. Model assessment and selection
    4.1. Generalization. Model complexity. Bias-variance tradeoff;
    4.2. Vapnik-Chervonenkis theory. Methods to reduce training error optimism – cross-validation and bootstrap;
5. Inference and model interpretation.
    5.1. Bootstrap and maximum likelihood methods.
    5.2. EM algorithm.
    5.3. Interpretation of linear regression and classification.
    5.4. GAM partial function interpretation. Odds ratio function interpretation.
    5.5. Extraction of model explanations – LIME (Local Interpretable Model-Agnostic Explanations).
6. Unsupervised learning
    6.1. Review of the concepts of cluster analysis and principal component analysis (PCA).
    6.2. Independent component analysis (ICA).
    6.3. Self Organizing Maps (SOM).

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

Os conteúdos programáticos abordados são fundamentais para que o aluno compreenda o processo de implementação de um modelo, sua aplicação e interpretação sob o ponto de vista estatístico, o que se encontra em coerência com os objetivos definidos.

Assim, os objetivos 1,2,3, relacionados com uma abordagem teórica do processo de modelação, são satisfeitos pelas seções 1,2,3 e 6. do programa. O objetivo da avaliação e seleção do modelo (objetivo 4) é alcançado pela seção 4 do plano de estudos. Em relação ao tema sobre interpretabilidade do modelo, definido no objetivo 5, este é satisfeito no ponto 5 do programa da unidade curricular. O objetivo 6, referente à implementação prática e computacional dos métodos utilizados para a construção e aplicação dos modelos, será satisfeito em todos os pontos do programa, uma vez que ambas as perspetivas, teórica e prática, serão abordadas simultaneamente em cada tema.

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

The syllabus is essential for understanding all the model construction process, its application, and interpretation from a statistical point of view, which is in line with the defined goals.

Thus, goals 1,2,3, which are related to a theoretical approach to the modeling process, are satisfied by the

program sections 1,2,3 and 6. Model assessment and selection (goal 4), is achieved by section 4 of the syllabus. Concerning the theme about model interpretability, defined in objective 5, it is satisfied in point 5 of the program. Objective 6, regarding the practical and computational implementation of the methods used to implement the models, will be satisfied by all points of the program, since the two perspectives, theoretical and practical one, will be addressed simultaneously in each theme.

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

As aulas são teórico-práticas. A parte teórica é fortemente baseada no formalismo matemático que permite o desenvolvimento de capacidade crítica no processo de modelação. Os alunos serão familiarizados com as técnicas de aprendizagem estatística, onde se incluem a inferência e a interpretação dos modelos.

Na componente prática são implementados computacionalmente os modelos e procedimentos abordados na parte teórica, usando um *software* livre (preferencialmente o R), com base em casos que podem ser reais ou simulados. São disponibilizados aos alunos elementos de apoio aos conteúdos programáticos.

A avaliação de conhecimentos compreende ambas as componentes teórica e prática. A parte teórica é constituída por um exame (nota mínima de 9,5 valores). A parte prática compreende um projeto de grupo com apresentação e discussão obrigatória (nota mínima de 9,5 valores). Os grupos a serem formados deverão ser constituídos no máximo por 5 elementos.

A nota final do aluno, NF, será obtida através da fórmula: NF=0,5 NT+0,5 NP, onde NT representa 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 strongly based on the mathematical formalism that allows the development of critical capacity in the modeling process. Students will be familiarized with the techniques of statistical learning, including inference and model interpretation.

In the practical component, the 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 to students.

Knowledge assessment comprises both components (theoretical and practical). The theoretical part consists of an exam (minimum score of 9.5). The practical part comprises a group's project with mandatory presentation and discussion (minimum score of 9.5). The groups should have a maximum number of 5 elements.

The student's final grade, NF, will be obtained using the formula: NF = 0.5 NT + 0.5 NP, where NT represents 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 modelação, principalmente naqueles que permitem a interpretação das estimativas e das associações entre variáveis. 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. Adicionalmente, na parte prática, o aluno pode aplicar as aptidões já mencionadas para resolver problemas reais ou simulados (próximos da realidade). O destaque dado às questões relacionadas com a interpretabilidade e inferência resultará num entendimento holístico do problema a resolver, o que será bastante útil em situações reais da futura vida profissional do aluno.

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 modeling process, especially those that allow the interpretation of estimates and associations between variables. It is important that the student acquires skills in the use of statistical reasoning to solve modeling 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 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. Hastie, T., Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer
2. Steyerberg, E. W. (2009). Clinical prediction models: A practical approach to development, validation, and updating. New York: Springer
3. Bishop, C. M. (2016). Pattern Recognition and Machine Learning.
4. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data mining: Practical machine learning tools and techniques.
5. Amaral Turkman, M.A. e Silva, G. (2000). Modelos Lineares Generalizados - da Teoria á Prática. Edições SPE,Lisboa.
6. Hastie, T. e Tibshirani, R. (1990). Generalized additive models. CRC Monographs on Statistics & Applied Probability. Chapman & Hall/CRC.
7. Efron, B. e Tibshirani, R. (1994). An Introduction to the Bootstrap. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.
8. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Muller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*.
9. Cadarso-Suárez, C., Roca-Pardiñas, J., Figueiras, A., & González-Manteiga, W. (April 30, 2005). Non-parametric estimation of the odds ratios for continuous exposures using generalized additive models with an unknown link function. Statistics in Medicine, 24, 8, 1169-1184.
10. Brás-Geraldes, C., Papoila, A., & Xufre, P. (April 01, 2020). Odds ratio function estimation using a generalized additive neural network. Neural Computing and Applications, 32, 8, 3459-3474.

[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.