UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Data Science (CDA)

33420 - Ciència de Dades

2020-2021

- José Hernández Orallo, DSIC, UPV, jorallo@upv.es
- Fernando Martínez-Plumed, DSIC, UPV, fmartinez@dsic.upv.es

Máster Oficial Universitario en
Ingeniería Informática
muiinf.webs.upv.es

MUIInf

Description

- Credits: 6.0 (1.5: theory, 3: seminar, 1.5: lab)
  - Theory and seminar will be intertwined.
- Lecturers
  - José Hernández Orallo (jorallo@upv.es)
    - Office 236, 2nd floor DSIC (Bldg. 1F).
    - Attention/tutoring hours: on demand by email.
  - Fernando Martínez-Plumed (fmartinez@dsic.upv.es)
    - Office 308, 3nd floor DSIC (Bldg. 1F).
    - Attention/tutoring hours: on demand by email.

Learning Outcomes

After completion of the course, the student will be able to understand the role of the data scientist in organisations, identify problems and opportunities and deploy solutions using off-the-shelf tools.

- Goals:
  1. recognise the value of data and the business opportunities for the development of data-driven products.
  2. determine the technologies that are needed to handle data efficiently in different environments, different sizes and formats, in order to ease data understanding and analysis.
  3. estimate the complexity and resources that are needed for a data analysis project and establish the measures of cost and success.

Learning Outcomes

- Specific objectives:
  - Realise the value of data and data-driven products.
  - Know the process of converting data into knowledge.
  - Use tools to integrate, prepare and visualise data.
  - Use a data-analysis language or tool to obtain models.
  - Evaluate models.
  - Deploy and exploit knowledge.

Contents

- **Unit 1**: Introduction (4,5h)
  - Data science: the role of the data scientist.
  - The value of the data: examples.
  - The D2K (Data to Knowledge) process.
  - Big Data: challenges and solutions.
- **Unit 2**: Data integration and manipulation (15h)
  - Source types and data repositories.
  - Data gathering, integration and cleansing
  - Data property, privacy and security.
  - Data visualisation and comprehension.
- **Unit 3**: Data analysis (17h)
  - Predictive and descriptive tasks
  - Supervised techniques
  - Non-supervised techniques
  - Model evaluation.
- **Unit 4**: Knowledge exploitation (5,5h)
  - Assistants, prescriptors and recommenders
  - Integration into decision making, dashboards and monitoring.

PLUS:

Introduction to **R** (5,5h)
Introduction to **Py**thon (3,5h)
Project Feedback, pre-**Ev**aluation (4,5h)
Final **Ev**aluation (4,5h)

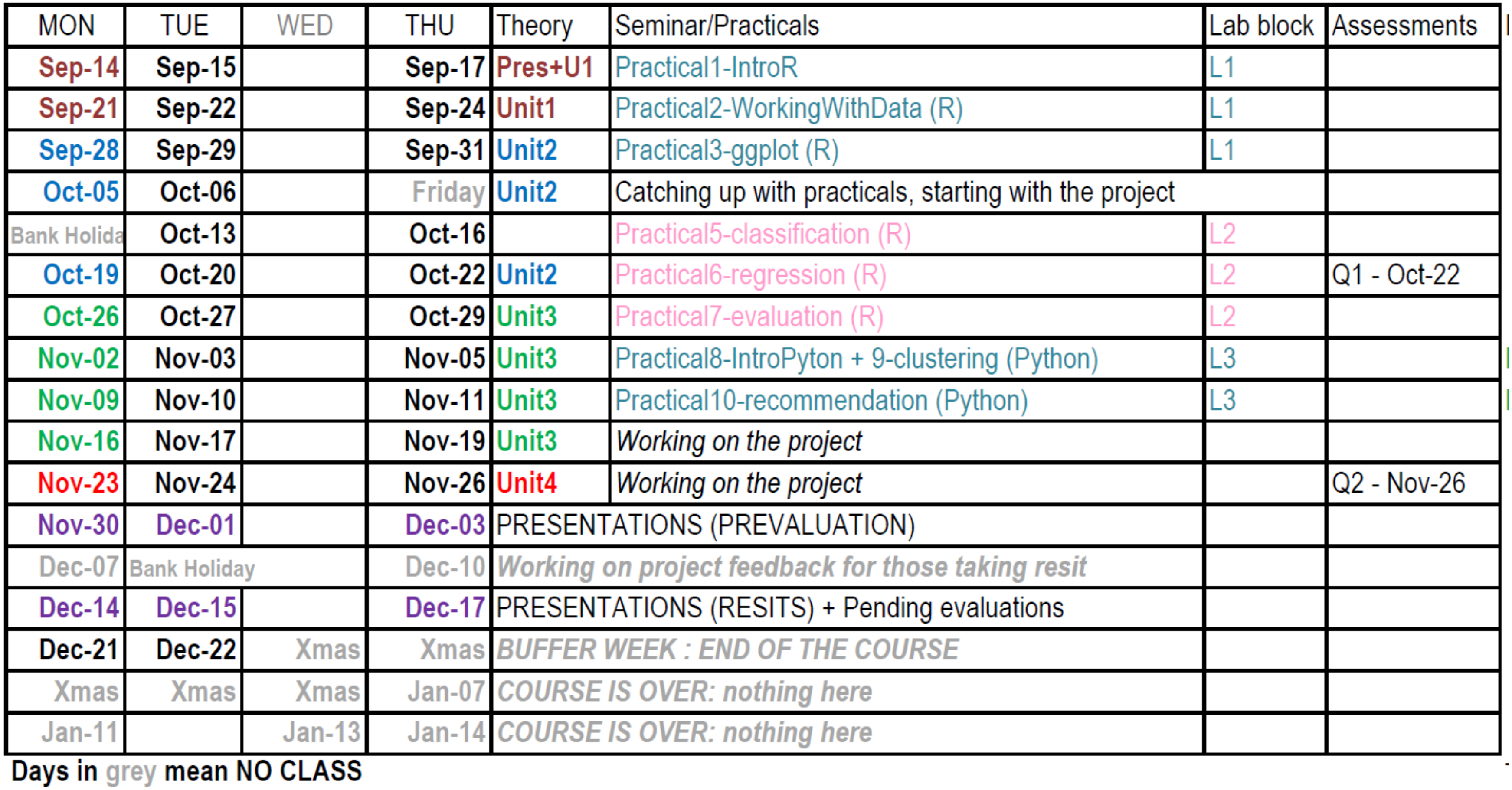
Evaluation

- Mostly practical evaluation:
  - Short questionnaires **in the classroom** (2): Q1, Q2 (10% each)
  - Short practical assignments (3): L1, L2, L3 (10% each)
    - Portfolio **delivered** on Poliforma't (assignment for each of the 10 practicals). Can be done in couples, but **evaluated individually through interview** at most 3 weeks after the start of that practical.
  - Freelance data scientist project: G1 (50%) * C1 (0-1)
    - Groups of **three students**.
    - Develop the idea of a new product from the use of data (open data, Internet, repositories, etc.) or that could improve an existing procedure with data-acquired knowledge.
    - **Oral presentation (pre-evaluation and final evaluation weeks)**.
    - **Evaluation rubric (G1)**: data value, alternatives and innovation, technical tool integration, project effort and exposition quality.
    - **Co-evaluation rubric (C1)**: percentage of contribution, disposition
    - Presentation **delivered** on Poliforma't.

Course plan

- Mon: 10:00:11:00 (Teams, recorded), Tue: 16:30-18:00 (Teams), Thu: 15:00-17:00 (1G 0.2)

  Except first 2 weeks, where everything will be on teams that week



MON	TUE	WED	THU	Theory	Seminar/Practicals	Lab block	Assessments
Sep-14	Sep-15		Sep-17	Pres+U1	Practical1-IntroR	L1	
Sep-21	Sep-22		Sep-24	Unit1	Practical2-WorkingWithData (R)	L1	
Sep-28	Sep-29		Sep-31	Unit2	Practical3-ggplot (R)	L1	
Oct-05	Oct-06		Friday	Unit2	Catching up with practicals, starting with the project		
Bank Holida	Oct-13		Oct-16		Practical5-classification (R)	L2	
Oct-19	Oct-20		Oct-22	Unit2	Practical6-regression (R)	L2	Q1 - Oct-22
Oct-26	Oct-27		Oct-29	Unit3	Practical7-evaluation (R)	L2	
Nov-02	Nov-03		Nov-05	Unit3	Practical8-IntroPyton + 9-clustering (Python)	L3	
Nov-09	Nov-10		Nov-11	Unit3	Practical10-recommendation (Python)	L3	
Nov-16	Nov-17		Nov-19	Unit3	*Working on the project*		
Nov-23	Nov-24		Nov-26	Unit4	*Working on the project*		Q2 - Nov-26
Nov-30	Dec-01		Dec-03	PRESENTATIONS (PREVALUATION)			
Dec-07	Bank Holiday		Dec-10	***Working on project feedback for those taking resit***			
Dec-14	Dec-15		Dec-17	PRESENTATIONS (RESITS) + Pending evaluations			
Dec-21	Dec-22	Xmas	Xmas	***BUFFER WEEK : END OF THE COURSE***			
Xmas	Xmas	Xmas	Jan-07	***COURSE IS OVER: nothing here***			
Jan-11		Jan-13	Jan-14	***COURSE IS OVER: nothing here***			

Days in grey mean NO CLASS

Recommended Readings

- Theory:
  - Foster Provost and Tom Fawcett "Data Science for Business: Fundamental principles of data mining and data analytic thinking", O'Reilly Media, 2013
  - Jeffrey Stanton "Introduction to Data Science", 2012.
    https://storage2.ischool.syr.edu/media.ischool.syr.edu/oldmedia/documents/2012/3/DataScienceBook1_1.pdf
  - Lars Nielsen, Noreen Burlingame "A simple introduction to data science", 2013 (ultra-short introduction)
  - Emmanuel Ameisen "Building Machine Learning Powered Applications", O'Reilly, 2020, https://www.oreilly.com/library/view/building-machine-learning/9781492045106/
  - Rachel Schutt "Doing data science", O'Reilly 2013
  - Jiawei Han "Data Mining: Concepts and Techniques", 3rd edition 2012.
  - Kirill Dubovikov "Managing Data Science: Effective strategies to manage data science projects and build a sustainable team", Packt Publishing, 2019
  - José Hernández-Orallo, M.José Ramírez-Quintana, Cèsar Ferri, "Introducción a la minería de datos", Pearson 2004
  - Peter Flach "Machine learning: the art and science of algorithms that make sense of data", Cambridge 2013.
- Lab (R and Python):
  - CRAN manuals: http://cran.r-project.org/doc/manuals/R-intro.pdf 2020, http://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf, 2014.
  - Luis Torgo "Data Mining with R", CRC Press 2010.
  - Wikibooks: http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R,2019
    http://en.wikibooks.org/wiki/R_Programming, 2019.
  - Graham Williams: Hands-On Data Science with R, http://onepager.togaware.com/
  - Wes McKinney "Python for Data Analysis Data Wrangling with Pandas, NumPy, and Ipython"
  - Toby Segaran "Programming Collective Intelligence: Building Smart Web 2.0 Applications", 2007
  - Raúl Garreta, Guillermo Moncecchi "Learning scikit-learn: Machine Learning in Python" 2013