Statistical Learning (2023)

Santiago Marco (Santiago.marco@ub.edu)
Department of Electronics and Biomedical Engineering, UB
Institute for Bioengineering of Catalonia (IBEC)

Oscar Lao (oscar.lao@ibe.upf-csic.es)
Institute for Evolutive Biology (UPF-CSIC)

Aims & Course Effort

- **Aims:**
  - By the end of the course the student will be able to develop statistical learning models from example datasets using state of the art methods and interpret correctly the model outputs.

- **4 ECTS: 4x30 = 120 hours dedication**
- **36 hours in classroom:**
  - 20 hours theory
  - 16 hours computational lab (in R)
- **2 hours Small Project Presentation**
- **4 hours exams (2+2)**
- **80 hours outside the classroom:**
  - 28 hours of independent study
  - 16 hours Small Project
  - 12 hours of directed group work (Lab reports & Questionnaires)
  - 4 hours of Reading
  - 20 hours R programming

DataCamp assignments (pending)

- **Intermediate R (6 hours)**
- **Introduction to Machine Learning (6 hours)**
  - What is Machine Learning
  - Performance Measures
  - Classification
  - Regression
  - Clustering
- **Unsupervised Learning in R (4 hours)**
  - Unsupervised Learning in R
  - Hierarchical Clustering
  - Dimensionality Reduction with PCA
  - Case study
- **Supervised Learning in R (4 hours)**
  - K-NN
  - Naïve Bayes
  - Logistic Regression
  - Classification Trees.

Labs (S. Marco) & Reading

- **Labs:**
  - Basic Classifiers
  - Classification mass spectrometry data (Basic Feature Extraction / Validation)
  - Classificaition metabolomics
  - Multilinear Regression

- **In Aula you will find a guide on how to write the lab report.**

- **Reading & Questionnaire:**

- **F. Rohart *et al.*, "mixOmics: An R package for 'omics feature selection and multiple data integration", PLOS Computational Biology, (2017)**

Course Contents & Key Dates

- W1: Introduction to Statistical Learning Theory (SM)
- W2: Introduction to Statistical Learning Lab (SM)
- W3: Feature Extraction Basics (SM)
- W4: Feature Selection (OL)
- W5: Custering (OL)
- W6: Statistical Classifiers – CrossValidation (SM)
- W7 : Partial Exam (29 Oct) - Adv. Classifiers (OL)
- W8: Adv. Classifiers (OL) –Multi Linear Regression (SM)
- W9: Multi linear Regression
- W10: Non-linear Regression

- **Final Exam: 5 Dec**
- Re-evaluation: 10 Jan

Evaluation

- **Partial Exam: 20%**
- **Final Exam: 20%**
- **Small Project: 30% (in teams of 4-5)**
- **Reports: 20%**
  - Questionnaires will be answered in the last 30 min of the lab. A mínimum of 8/10 is needed for the reports to be considered for the final mark,
- **Questionnaires/Exercises: 10%**

- **Exams are based on a questionnaire and the solutions of practical exercises in R.**
- **A minimum mark of 4.5/10 as the average of the partial and final exam is needed to consider the rest of the activities.**
- **Attendance to the labs is mandatory**

- **The Small Project will be assessed based on an oral presentation and the submission of the data analysis code. No report will be needed.**

Literature

- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
- **Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. 2n edition. New York: Springer series in statistics, 2017**.
- C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- Wehrens, Ron. *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springer Science & Business Media, 2011.
- Varmuza, Kurt, and Peter Filzmoser. *Introduction to multivariate statistical analysis in chemometrics.* CRC press, 2016.
- Witten Ian H, Frank Eibe and Hall Mark A. *Data Mining*. Morgan Kaufmann.
- Haupt and Haupt. *Practical genetic algorithms*. John Wiley & Sons.
- Cox, Trevor and Cox, Michael. *Multidimensional Scaling*. CHAPMAN & HALL/CRC
- Greenacre. *Correspondence analysis in practice*. CRC
- Hair, Anderson, Tatham, Black. *Analisis multivariante*. Prentice Hall.

Before we start (OL)

- **Some basic rules in my class**
  - Each session is divided in two parts of ~50 minutes.
  - The door closes 10 minutes after we start the session.
- **Classroom dynamic**
  - Introduction of a practical problem
  - Brief description of a theory to solve the problem (first hour)
    - A lots of questions to be expected!
      - The only wrong answer will be not giving any answer!
  - Implement the algorithms / pipeline using an R package in R (second hour) = TOY EXAMPLES
  - Solve the problem and provide a short report

Before we start (OL)

- **Short report (Parts to be filled)**
  - What is the problem you had to address & How did you address it?
    - Short introduction (half page or so) of what is the question to answer, and a brief description of the techniques you applied. Any source is welcomed.
    - HOWEVER, PLEASE DO NOT COPY PASTE without citing where does it comes = **PLAGIARISM**!): "Put in your own words everything, so I can see that you UNDERSTAND what you are doing!"
  - Methods
    - R code that you used, LINE BY LINE MEANINGFULY COMMENTED!
  - Description of the results. Figures, Tables, Statistics of the performance
  - Interpretation of the results
- Introduction + Results + Interpretation ~ two pages