UNIVERSITÀ DEGLI STUDI
DI TRENTO

Dipartimento di Dipartimento di Matematica
Academic year 2019/2020

Data analysis and exploration [145707]

No class division

Study course Data Science
Regulation Data Science
Curriculum standard

Lecturers: MARIO LAURIA (Tit.)

Hours amount: 48

Period: Second semester

Credits: 6

Fields: INF/01

Formative aims
The goal of this course is to enable students to analyze various types of high-dimensional data types commonly encountered in the practice of Molecular Biology using a range of methodologies, including established statistical approaches, machine learning, and recently introduced methods.
At the end of the course, students are expected to possess the following skills:
- familiarity with different types of Molecular Biology data and their specificity;
- good understanding of different conceptual methods for analyzing the data;
- familiarity with practical computational tools for carrying out the analysis;
- ability to frame the results of their analysis in a biological context using networks;
- ability to provide a functional interpretation of the results of their analysis;
- ability to complete a project, write a structured report, and orally present the results of their work.

Prerequisites
This course requires familiarity with principles taught in the introductory Math and Computer Science courses normally offered in Laurea Triennale/Bachelor curricula, particularly Analysis/ Calculus, Geometry, Computer Programming. Some introductory course in Statistics and in Molecular Biology are highly recommended. The content of the course will be calibrated according to the student backgrounds, possibly with the addition of short reviews or introductory sessions as needed to reach the minimum level of competence necessary for the class.

Course programme
Introduction to data types in Molecular Biology
Biological data processing in R
Introduction to explorative analysis using PCA
Introduction to Machine Learning: classification, clustering, supervised/unsupervised learning
Introduction to classification using LDA
Regularized regression: Lasso/Ridge regression
Rank-based signatures
Biological networks and their visualization
Network analysis with applications to biological networks
Functional analysis of biological data
Final review and projects discussion

Teaching methods
Traditional lectures will be accompanied by weekly practice sessions in which the newly learned methods will be applied using R. To complete the course students are required to carry out a semester long project. As part of the project, students will select a real data set of their choice on which to apply week by week all the methods learned

during the class. The weekly practice sessions will be employed by each student to derive the results that collectively will form the final report. This way over the course of the semester students will be able to prepare their project mostly during the class hours, with assistance from the instructor if needed.

Learning assessment method
As final exam each student will be required to give a succinct oral presentation of his/her project, including description of the selected data set, of the methods of analysis and notable results. Alternatively, in exceptional and well motivated cases a traditional oral exam can be requested covering all the practical and theoretical aspects of the course.

Reference books
The course has been designed to be self-contained, and all the course material, including R code, will be made available online. A short selection of textbooks for different parts of the course is reported below.

- Julian J. Faraway, Practical Regression and Anova using R. available at
http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf
- Venables W.N, Ripley B.D. Modern Applied Statistics with S-PLUS. Springer, New York, 1997.
- Flury B. A first course in multivariate statistics. Springer, New York, 1998.
- Heiberger, R.M. and Holland,B. Statistical analysis and data display, Springer, New York, 2004

Further information
none

Printed on 25/06/2020