האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

The Hebrew University of Jerusalem

Syllabus

Statistical learning for data science - 52016

Last update 20-12-2023

HU Credits: 4

Degree/Cycle: 2nd degree (Master)

Responsible Department: Statistics

Academic year: 0

Semester: 2nd Semester

Teaching Languages: Hebrew

Campus: Mt. Scopus

Course/Module Coordinator: Barak Sober

Coordinator Email: Barak.Sober@mail.huji.ac.il

Coordinator Office Hours:

Teaching Staff:

Dr. Barak Sober,
Mr. Gidi Yoffe

Course/Module description:

The course deals with the statistical analysis of large modern datasets. We will discuss statistical and computational challenges, and learn general principles as well as specific methods of analysis.

In particular, we focus on exploratory data analysis, and on prediction models.

Labs (hand-in assignments) are an important part of the course. Students will analyze real data sets. They will also compare methods on real and simulated data.
Learning Outcome:
Students will be able to apply a wide range of methods in context of supervised and non-supervised learning as well as to assess the quality of the outcome. In addition, the students will acquire some theoretical understanding of the methods, be able to discern which method apply to which situation, and understand the theoretical limitations of the various methods.

Course/Module aims:

The course aims to present modern data analysis techniques. The goal is also for the students to learn and practice research work in data analysis and statistical method development.

Learning outcomes - On successful completion of this module, students should be able to:

At the end of this course, students will be able to:
- Examine a data set and display its features
- Postulate a research interest as a prediction problem, and understand the advantages and disadvantages of the prediction paradigm compared to other types of inference
- Construct a prediction model (categorical or continuous)
- Quantify the success of the model, and compare different methods or models. Estimate the error and uncertainty.
- Communicate the analysis in writing.

Attendance requirements(%):

80

Teaching arrangement and method of instruction: The course will be frontal in the classroom (a few lessons will be given through Zoom).
Attendance is mandatory.
The lessons will be recorded but will become available only to students that will show a legitimate reason for their absence.

Course/Module Content:
The list below is just a tentative list of topics. The actual topics may differ (some may be omitted while others added).
1. Cleaning and exploring data
2. PCA
3. Representation and distances
4. Clustering
5. Stability and Bootstrap
6. Introduction to supervised learning, bias vs. variance
7. Regression: expanding basis + wavelets
8. Regularized regression: Ridge, Lasso, Elastic Net
9. Regression trees
10. Support Vector Machines and Reproducing Kernels
11. Discriminative analysis
12. Boosting
13. Intro to Neural Network

Required Reading:
None

Additional Reading Material:
Understanding Machine Learning (Shalev-Shwartz & Ben-David)
https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

A Distribution-Free Theory of Nonparametric Regression (Gyorfi, Kohler, Krzyzak & Walk)
https://web.stanford.edu/class/ee378a/books/book1.pdf

Grading Scheme:
Essay / Project / Final Assignment / Home Exam / Referat 75 %
Submission assignments during the semester: Exercises / Essays / Audits / Reports / Forum / Simulation / others 25 %

Additional information: