The Hebrew University of Jerusalem

Syllabus

Topics in high dimension probability with application to data science - 52027

Last update 30-08-2022

HU Credits: 3

Degree/Cycle: 1st degree (Bachelor)

Responsible Department: Statistics

Academic year: 0

Semester: 1st Semester

Teaching Languages: Hebrew

Campus: Mt. Scopus

Course/Module Coordinator: Ariel Jaffe

Coordinator Email: ariel.jaffe@mail.huji.ac.il

Coordinator Office Hours: Monday 16:00

Teaching Staff:

Dr. Ariel Jaffe

Course/Module description:

Recent years have witnessed a dramatic increase in the dimension and complexity of datasets acquired in many scientific domains. Inference from high-dimensional observations poses new methodological and theoretical challenges for researchers in statistics and data science. In this course, our main objective is to provide the theoretical foundations required for deeper understanding of important data science challenges, such as dimensionality reduction, clustering and covariance estimation, low-rank matrix completion, and sparse recovery. The course lectures will derive key results from topics in high-dimensional probability and random matrix theory and apply them for specific examples and applications. The course has three main parts: (i) Background on tail bounds, the law of large numbers, concentration, and sub-Gaussian and subexponential random variables. (ii) Random vectors in high dimensions, including the concentration of the norm, uniform concentration bounds, and the Johnston Lindenstrauss lemma, and (iii) random matrices, including the norm of random matrices and the matrix Bernstein inequality.

Course/Module aims:

Provide theoretical tools for deeper understanding and analysis of various applications in data science.
During the course, the students will be familiarized with multiple high-dimensional probability techniques and bounds and gain experience in applying them to derive finite-sample guarantees for important applications in supervised and unsupervised learning.

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

- *The student will learn to apply a variety of techniques for proving tail inequalities.*
- *The student will apply tail inequalities to provide theoretical support for various data science applications.*
- *The student will simulate various settings to compare numerical results with*

theoretical expectations.

Attendance requirements(%):

Teaching arrangement and method of instruction: Lectures

Course/Module Content:

Course content (partial)

1. Part I: Scalar concentration and tail bounds

- Types of convergence, basic tail bounds, convexity
- Laws of large numbers, the Berry-Essen inequality and the Delta method.
- Basic concentration inequalities: Hoeffding and Chernoff.

Applications: boosting and degrees of random graphs.

2. Part II: High dimensional vectors

- Concentration of norm of random vector, concentration of Lipschiz function of random vector
- The Johnson–Lindenstrauss lemma
- application: dimensionality reduction

3. Part III: Random matrices

- Assymptotic results: The Marchenko-Pastur and semi-circle law
- Concentration of norm of random matrix
- The matrix Bernstein inequality

Application: covariance estimation

- Matrix Chernoff inequality

application: Singular values of submatrices
application: Connectivity of random-graphs

Required Reading:

[1] Joel A Tropp et al. “An introduction to matrix concentration inequalities”. In: Foundations and Trends® in
Machine Learning 8.1-2 (2015), pp. 1–230.

[2] Roman Vershynin. High-dimensional probability: An introduction with applications in data science. Vol. 47.
Cambridge university press, 2018.

[3] Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. Vol. 48. Cambridge University
Press, 2019.
2

Additional Reading Material:

Grading Scheme:

Additional information: