האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

The Hebrew University of Jerusalem

Syllabus

Advanced topics in data analysis - 98449

Last update 14-02-2019

HU Credits: 2

Degree/Cycle: 2nd degree (Master)

Responsible Department: Public Health

Academic year: 0

Semester: 2nd Semester

Teaching Languages: Hebrew

Campus: Ein Karem

Course/Module Coordinator: Dr. Shai Carmi

Coordinator Email: shai.carmi@huji.ac.il

Coordinator Office Hours: Please coordinate by email

Teaching Staff:

Dr. Shai Carmi
Ms.

Course/Module description:
The class will cover modern topics in statistics relevant for biomedical and epidemiological research.
The material will include hypothesis testing, estimation, and classification. Particularly, we will consider studies of large-scale, high dimensional datasets ("big data"), as well as studies with a relatively small number of observations. The class will highlight the opportunities in data analysis, and review advantages and disadvantages of the various methods covered.
Problem sets will mostly include programming exercises in the R language.

Course/Module aims:

Learning outcomes - On successful completion of this module, students should be able to:
Graduating students should be able to:
** Determine the most appropriate and powerful statistical test for their own data analysis problems, analyze the data and interpret the and results.*
** Recognize the problem of multiple comparisons and control the false discovery rate.*
** Understand when resampling methods (permutations, bootstrap) are useful for hypothesis testing and estimation and be able to apply them.*
** Understand the goals and theoretical basis of machine learning.*
** Apply common methods for classification and understand their properties.*
** Apply common methods for clustering and dimension reduction.*
** Implement the newly studied methods in the R programming language.*

Attendance requirements(%):

Teaching arrangement and method of instruction: Lectures

Course/Module Content:
1) An introduction to statistical inference: hypothesis testing and estimation, the sampling distribution, the confidence interval, errors and power, bias and variance.
2) Robustness of statistical tests: what common tests assume and how to determine

whether the assumptions hold.
3) What to do in case assumptions don't hold: transformations, non-parametric tests (sign test, Wilcoxon test, confidence interval for the median).
4) Likelihood-based methods: maximum likelihood estimation, the likelihood ratio test, properties.
5) Multiple testing: motivation, Bonferroni correction and its properties, the false discovery rate, the Benjamini-Hochberg method.
6) Resampling methods for hypothesis testing and estimation: drawing random numbers from a distribution, the bootstrap, permutation tests.
7) Introduction to machine learning: goals and examples, types of algorithms, bias and variance, overfitting and its importance, cross-validation.
8) Classification: the kNN method, the perceptron and neural networks, (time permitting: decision trees). Advantages and pitfalls of common methods, measures of accuracy.
9) Unsupervised learning: clustering: k-means and UPGMA, dimension reduction using PCA.
10) Basic programming in R and applications to the material studied.

Required Reading:
None

Additional Reading Material:
An Introduction to Statistical Learning: with Applications in R, Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Springer, 2013.
Learning From Data, Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. AMLBook, 2012.
Pattern Recognition and Machine Learning, Christopher M. Bishop, Springer, 2007.
All of Statistics: A Concise Course in Statistical Inference, Larry Wasserman, Springer, 2004.

Course/Module evaluation:
End of year written/oral examination 70 %
Presentation 0 %
Participation in Tutorials 0 %
Project work 0 %
Assignments 30 %
Reports 0 %
Research project 0 %
Quizzes 0 %
Other 0 %

Additional information:

The class requires basic knowledge in statistics, as provided to first year students of medicine, biomedical sciences, or public health. Please refresh your memory with the principles of basic statistic before the beginning of the semester.

The class will emphasize the intuition behind the different methods and how to practically apply them, avoiding unnecessary mathematical details.

Knowledge of at least one programming language is a prerequisite to the class. The recommended language is R, but similar tools could also be used, such as Python or Matlab.

Problem sets will consist mostly of programming exercises. The exercises may be challenging and require significant effort for students with programming experience.