Statistics for Big data

Dimitris Karlis, Professor, AUEB, karlis@aueb.gr

Overview

The era of big data has led to paradigm shift in statistical methods. While statistics is an important pylon of data science, new challenges and new problems have occurred due to the abundance of data. The aim of the present course if to discuss the changes on the statistical methods and how the huge volume of data affects the classical statistical methods, while at the same time, how the new problems occurring can be solved via state of the art new statistical methodologies. We have selected some methods and problems to exploit showing the new potential and the new dynamics of statistical science towards new problems.

Key Outcomes

After completing the course, the students will be able to:

- understand the new challenges and problems due to the abundance of data
- apply new techniques defined to handle problems with big data
- implement the new methods using R

Requirements and Prerequisites

This course combines theory and practice, including multiple individual exercises and a final project using some of the methods discussed with real data. The course does not assume any prior experience in R, however, basic knowledge of programming and computer science concepts is required. The course implies a good prior knowledge of statistics.

Required Course Materials

There is no required textbook. All course materials will be provided in class and available for downloading.

Books

There are many books on the subject; the following selection provides a good foundation for those students who wish to delve deeper on the topics discussed in class:

- *Christophe Giraud (2015). Introduction to High-Dimensional Statistics. Philadelphia: Chapman and Hall/CRC.*

- *T. Tony Cai, Xiaotong Shen, ed. (2011). High-dimensional data analysis. Frontiers of Statistics. Singapore: World Scientific.*
- *Peter Bühlmann and Sara van de Geer (2011). Statistics for high-dimensional data: methods, theory and applications. Heidelberg; New York: Springer.*
- *T. Hastie, R. Tibshirani and R. Friedman (2009) Elements of Statistical Learning, Springer.*
- *E. D. Kolaczyk (2014) Statistical Analysis of Network Data with R. Springer*

Software/Computing requirements

- R free software environment for statistical computing and graphics. https://www.r-project.org/. Special packages will be used. Students are required to download an up-to-date version of R.

Grading

Students will be graded as follows:

- Participation - 10% (individual assessment).
- Mini assignments - 20%. These are individual assignments, such as simple exercises, small essays, work with data, search for cases, etc. that relate to each lecture and aim at establishing the student's comprehension of the lecture. They have to be submitted after each lecture.
- Group - Project - 60%. There will be one group project. Groups will be randomly formed. The group project includes the submission of documentation in the form of a detailed report.
- Final presentation – 10%.   It refers to the presentation of the group projects above. Here the interest lies on the personal oral communication of the findings/ methodology and the entire story in a scientifically correct and fascinating manner.

The course does not have exams.

Participation

In-class contribution accounts to a 10% of your grade and is an important part of our shared learning experience. Your active participation helps us to evaluate your overall performance. You can excel in this area if you come to class on time and contribute to the course by:

- Providing strong evidence of having thought through the material.
- Advancing the discussion by contributing insightful comments and questions.
- Listening attentively in class.
- Demonstrating interest in your peers' comments, questions, and presentations.
- Giving constructive feedback to your peers when appropriate.

Please arrive to class on time and stay to the end of the class period. Chronically arriving late or leaving class early is unprofessional and disruptive to the entire class. Repeated tardiness will have an impact on your grade. *Turn off all electronic devices prior to the start of class. Cell phones, tablets and other electronic devices are a distraction to everyone.*

Assignments

Late assignments will either not be accepted or will incur a grade penalty unless due to documented serious illness or family emergency. Exceptions to this policy for reasons of civic obligations will only be made available when the assignment cannot reasonably be completed prior to the due date, you make suitable arrangements, and give notice for late submission in advance.

Attendance Requirements

Class attendance is essential to succeed in this course and is part of your grade. An excused absence can only be granted in cases of serious illness or grave family emergencies and must be documented. Job interviews and incompatible travel plans are considered unexcused absences. Where possible, please notify the instructor in advance of an excused absence.

Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives explicit permission.

Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one's own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Please report observed violations of this policy. Any violations will incur a fail grade at the course and reporting to the senate for further disciplinary action.

Course Syllabus

The course comprises **six** units of three hours each. There will be one more session for presenting the projects.

Unit 1: Introduction – statistics in the era of big data

This lecture will discuss the problems appearing when trying to apply traditional methods to big datasets. Also new problems and challenges will be discussed, like variable selection, multiplicity, spurious correlations, sparse methods, computational issues and many others. The purpose is to understand the paradigm shift and introduce the new and fresh ideas.

Unit 2: Old methods new tricks

The purpose of this lecture is to show by examples problems appearing to apply well known methods like regression to big datasets. There will be a discussion of the kind of problems and how we can avoid/overcome them using new approaches. Examples using regression methods, clustering and classification will be discussed. The aim is to recognize the kind of problems and see state of the art recent approaches.

Units 3: Regularization methods

Variable selection becomes an important issue and problem when working with abundance of data and available variables. Traditional variables selection approaches are not easy to implement and new techniques have been proposed. This lecture will introduce the idea of regularization, explain the idea and show the relationship with other methods. Practical implementations for regression, GLM models and well as discriminant analysis will be exploited.

Units 4: Multiplicity problems and solutions.

A common problem when working with many variables relates to the multiplicity problem: i.e. we need to make a huge number of hypothesis testing and hence the probability of error increases dramatically. This lecture will discuss this problem and explain simple and more advance approaches to handle this. Issues like the False Discovery Rate, the Family- wise Error will be introduce and provide methods to control for multiple testing in several problems. Examples will be discussed.

Unit 5: Genetic Applications

In recent years we have seen a tremendous improvement on genomics, i.e. on research based on genes. Several methods have been developed for such data. A typical problem with such data is that we have a huge number of variables and thus special methods have been developed. The aim of this lecture is to introduce some methods suitable for such data and describe the relationship with existing methods.

Methods like SAM (significance analysis of Microarrays) which relates to randomization testing, Supervised PCA, Genome wide Association etc will be discussed.

Unit 6: Statistical Network Analysis

It is very common to have data represented as a network. This relates to social networks like FB and/or twitter or simpler cases like who send email to whom etc. Such data need special methods to work with. So in this lecture we will introduce some such methods for some special networks, including bipartite networks, latent space models, random subgraphs, etc. and showing how the relationships on such data can be modeled.