Statistical Learning and Big Data
Athens University of Economics & Business
Department of Economics
(MSc in Applied Economics and Finance & MSc in Finance and Banking)
Instructor: Aikaterini Kyriazidou

Course description:

This course is an introduction to statistical learning. It aims to familiarize students with the concepts of supervised and unsupervised learning for exploratory analysis and prediction using prominent techniques, such as regression, classification and clustering. In the process, some widely used techniques in data analysis will be introduced, such as principal component analysis, regularization, cross validation, and bootstrapping among others.

The course aims primarily to give a working knowledge and hands-on experience with the above tools and techniques, and therefore a part of it will also be dedicated to familiarization with using large databases (big data) and understanding the logic behind data structures, data handling, data manipulation and data quality assessment. To that end, an introduction to Structural Query Language (SQL) of relational databases will also be provided.

Statistical analysis and machine learning will be done using the R statistical computing package. Students will be required to do approximately six labs using real-world data. Extensive guidance in using R and SQL will be provided, but previous basic programming skills in R or exposure to a programming language such as MATLAB or Python will be useful.

Learning outcomes:

Upon successful completion of this module, the students should

- Have experience with the "art" of modeling and predicting real world phenomena
- Be able to perform regression, classification, clustering and principal component analysis using R
- Be able to handle complex databases using SQL
- Be able to understand, interpret, and critically evaluate data analysis performed by others
- Be familiar with R, a cutting-edge statistical software increasingly used in business, government and academic settings
- Be able to present and write a short report on a statistical learning problem using real data

Teaching and Learning Methodologies:

We will go over both theory and R programming in class and use whiteboard, slides and real-time code execution. I will mostly follow the required textbook *(An Introduction to Statistical Learning)*. Students are welcome to use the online resources available at the textbook site (http://www-bcf.usc.edu/~gareth/ISL/). TA sessions will mainly go over the textbook's R labs and SQL.

Technology:

You will need R (https://cran.r-project.org) and RStudio (https://www.rstudio.com) installed on your computer along with additional R packages (libraries), such as markdown, ISLR and knitr. The lectures and TA lab sessions will include applications using R. You will be asked to submit your homework's code in R Markdown, a simple formatting syntax for authoring HTML, PDF, and MS Word documents that contains chunks of embedded R code. We will mostly use the data sets accompanying the textbook and other publicly available data resources that will be indicated as we go. You may use your laptops during lectures only to browse lecture material, take notes and run code. Statistical calculators are not necessary. The use of mobile phones, including calculator apps, is not allowed during lectures.

Prerequisites:

- Basic matrix algebra and multivariate calculus
- Basic probability and distribution theory; statistical inference (point and interval estimation and testing)
- Regression analysis

Assessment:

- Homework - Project: 50%
- Final: 50%

Module 1: Introduction to Statistical Learning

The goal of this module is to introduce students to widely used statistical learning methods. These methods use a variety of computational tools for understanding large, complex datasets.

The following topics will be covered:

1. Introduction to Statistical Learning
2. Linear Regression
    - Estimation
    - Testing
    - Prediction
3. Classification Analysis
    - Linear Probability Model
    - Logistic and Probit Regression

    - Discriminant Analysis
4. Resampling Methods
    - Cross Validation
    - The Bootstrap
5. Model Selection and Regularization
    - Subset selection
    - Shrinkage Methods
6. Non-Linear and Nonparametric Methods
    - Polynomial Regression
    - Splines
    - Local Regression
    - K Nearest Neighbors
7. Tree-Based Methods
    - Decision Trees
    - Bagging, Random Forests, Boosting
8. Unsupervised Learning
    - Principal Components Analysis
    - Clustering (K-Means, Hierarchical Clustering)

Textbook

- *Introduction to Statistical Learning: with Applications in R*, By James, G., Witten, D., Hastie, T., Tibshirani, R. Springer, 2013, (available online at https://www-bcf.usc.edu/~gareth/ISL/)

Additional Recommended Resources

- *The Elements of Statistical Learning*, 2nd edition, by Trevor Hastie, Robert Tibshirani and Jerome Friedman
- https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/
- http://cran.r-project.org/doc/manuals/R-intro.pdf
- *R for Data Science*, by Garrett Grolemund and Hadley Wickham (https://r4ds.had.co.nz/index.html)
- Duke University, scientific writing resource https://cgi.duke.edu/web/sciwriting/

Module 2: Introduction to SQL

The aim of this module is to provide the programming tools for both R and SQL, with the view of combining them, so as to perform data handling and statistical/econometric analysis in an automated and systematic manner. For the purposes of the module, a student-accessible PostgreSQL Server will be deployed.

The following topics will be covered:

1. Introduction to SQL
2. Extracting Information
3. Data manipulation
4. Grouping and aggregating data
5. Selecting data from different tables

Textbooks

- *Beginning SQL*, Wilton Paul & Colby W. John, Wiley Publishing, Inc, 2005
- *Database Management Systems*, Ramakrishnan R. & Gehrke J., McGraw Hill, 2003, Third Edition.
- *Database Systems: The Complete Book*, Garcia-Molina H. & Ullman D. J. & Widom J., Prentice Hall, 2008, Second Edition

Additional Material

- *PostgreSQL 9.5.1 Documentation*, The PostgreSQL Global Development Group, 2016 (available online)
- *SQL Workbench/J User's Manual* (available online)