

Στατιστική Μεγάλων Διαστάσεων (High Dimensional Statistics)

Διδάσκων: Παναγιώτης Παπασταμούλης

Γενικά στοιχεία μαθήματος

Κωδικός:	61227
Τύπος:	Επιλογής
Επίπεδο:	Μεταπτυχιακό
Έτος σπουδών:	Β'
Εξάμηνο σπουδών:	4ο
ECTS:	5
Γλώσσα διδασκαλίας:	Ελληνική/Αγγλική

Course Description

Modern statistical applications deal with the analysis of high-dimensional datasets. However, inference is extremely challenging due to the curse of dimensionality: as the number of variables increases, the data may become quite noisy and it is difficult to separate the actual underlying information. The course will present modern statistical techniques particularly suited to problems with high dimensionality, such as:

- large scale hypothesis testing
- statistical applications in bioinformatics
- applications of (generalized) linear models in big data problems
- Regularization techniques (LASSO, Ridge regression)
- Classification and model-based clustering for multivariate data with dimensionality reduction techniques
- Statistical network analysis

Περιγραφή Μαθήματος

Σύγχρονες εφαρμογές της Στατιστικής απαιτούν τη διαχείριση και ανάλυση μεγάλου όγκου δεδομένων. Παρολαυτά, η αύξηση της διάστασης των δεδομένων δεν συνεπάγεται πάντα αύξηση της χρήσιμης πληροφορίας, αλλά και αύξηση του θορύβου που εμπεριέχεται σε αυτά. Έτσι, τίθενται νέες προκλήσεις στη Συμπερασματολογία. Το μάθημα της Στατιστικής Μεγάλων Διαστάσεων θα παρουσιάσει σύγχρονες Στατιστικές τεχνικές για την αντιμετώπιση τέτοιων προβλημάτων, όπως:

- Προβλήματα πολλαπλών συγκρισίων και έλεγχος του ρυθμού των ψευδώς θετικών αποτελεσμάτων
- Στατιστικές εφαρμογές σε προβλήματα βιοπληροφορικής
- Εφαρμογές (γενικευμένων) γραμμικών μοντέλων σε προβλήματα μεγάλων διαστάσεων
- Τεχνικές κανονικοποίησης (LASSO, Ridge regression)
- Διαχωριστική ανάλυση και ανάλυση συστάδων σε πολυμεταβλητά δεδομένα με τεχνικές μείωσης διάστασης
- Στατιστική ανάλυση δικτύων

Prerequisites

- Prerequisites
 - Hypothesis Testing - Linear Models
 - Generalized Linear Models
- Some basic knowledge on
 - Computational Statistics
 - **R** programming
- Students will also benefit from the following courses (not required)
 - Applied Bayesian Statistics
 - Statistical Learning

Προαπαιτούμενα

- Καλή γνώση
 - Εκτιμητικής και Ελέγχων Υποθέσεων
 - (Γενικευμένων) Γραμμικών Μοντέλων
- Βασικές γνώσεις
 - Υπολογιστικής Στατιστικής
 - Προγραμματισμού σε **R**
- Σχετικά προσφερόμενα μαθήματα (δεν είναι προαπαιτούμενα)
 - Εφαρμοσμένη Στατιστική κατά Bayes
 - Στατιστική Μάθηση

Key Outcomes

After completing the course, the students will

- know the challenges that big data impose to the statistician
- properly deal with large scale hypothesis testing
- put their hands on many different types of data from cutting edge research areas such as bioinformatics and network analysis
- learn many novel statistical ideas and methods developed in the last 20 years
- know how to apply the relevant methods using **R** and **Bioconductor**.

Επιδιωκόμενα μαθησιακά αποτελέσματα

Μετά την επιτυχή ολοκλήρωση του μαθήματος, οι φοιτητές θα είναι σε θέση

- να γνωρίζουν τις προκλήσεις που θέτουν τα δεδομένα μεγάλων διαστάσεων στη Στατιστική Συμπερασματολογία
- να αντιμετωπίζουν σωστά προβλήματα ελέγχου πολλαπλών υποθέσεων
- να έρθουν σε επαφή με σύνολα δεδομένων από σύγχρονες ερευνητικές περιοχές όπως η Βιοπληροφορική και η ανάλυση δικτύων
- να μάθουν νέες Στατιστικές μεθοδολογίες που αναπτύχθηκαν τα τελευταία 20 χρόνια ειδικά σχεδιασμένες για την αντιμετώπιση προβλημάτων μεγάλων διαστάσεων
- να υλοποιούν κατάλληλους αλγορίθμους στην **R** και **Bioconductor**.

Teaching and Learning Activities

One three-hour lecture per week, study exercises and programming exercises as homework (some to be submitted).

Διδακτικές και Μαθησιακές Μέθοδοι

Μια διάλεξη τριών ωρών εβδομαδιαίως, ασκήσεις μελέτης και προγραμματισμού στο σπίτι (ορισμένες προς παράδοση).

Assessment and Grading Methods

The final grade will be a weighted average of 2 or 3 homework assignments.

Μέθοδοι Αξιολόγησης και Βαθμολόγησης

Ο τελικός βαθμός θα προκύψει ως σταθμισμένος μέσος όρος 2 ή 3 εργασιών.

References

- [1] Christophe Giraud (2015). Introduction to High-Dimensional Statistics. Philadelphia: Chapman and Hall/CRC
- [2] Tony Cai, Xiaotong Shen, eds. 2011. High-dimensional data analysis. Frontiers of Statistics. Singapore: World Scientific
- [3] Hastie, R. Tibshirani and R. Friedman (2009). Elements of Statistical Learning. Springer
- [4] Efron, Bradley. Large scale inference: Empirical Bayes Methods for Estimation, Testing and Prediction. Cambridge University Press, 2010
- [5] McLachlan, G. and Peel, D (2000). Finite Mixture Models. Cambridge University Press, 2010
- [6] Wasserman, S. and G. Robins (2005). An introduction to random graphs, dependence graphs, and p^* . Models and methods in social network analysis.
- [7] Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. Journal of the American Statistical Association.