

קורסים ארציים בסטטיסטיקה לתלמידי תואר שני ושלישי, תשפ"ד

בהתאם להכרזת ראשי האוניברסיטאות, הקורסים הארציים בסטטיסטיקה לתלמידי תואר שני ושלישי יתקיימו במתכונת מצומצמת של 11 שבועות. להלן גרסה עדכנית של התוכנית לשנה זו.

הקורסים יינתנו בקמפוס של אוניברסיטת תל אביב אך לאור המצב יתקיימו באופן היברידי (פרונטלי + זום) והקלטות השיעורים יועלו לאתרי הקורסים. כל תלמידי המוסדות האקדמיים השונים מוזמנים להירשם ולהשתתף ללא תשלום נוסף, וציונם בקורסים יועבר למוסד האם.

התוכנית

סמסטר א, יום ה

9:10 -- 11:00 מס' הקורס 3654109
סדרות עיתיות בכלים של סטטיסטיקה ולימוד מכונה, פרופ' יאיר גולדברג, טכניון

12:10 -- 14:00 מס' הקורס 3654111
תיאוריה של למידה, ד"ר עמיחי פיינסקי, אוניברסיטת תל אביב

15:10 -- 18:00 מס' הקורס 3654078
סטטיסטיקה לעידן ה Big Data, פרופ' סהרון רוסט, אוניברסיטת תל אביב

מפגשים

4.1 11.1 18.1 25.1 1.2 8.2 15.2 22.2 29.2 7.3 14.3

הקורסים יינתנו בעברית, אלא אם תהיה דרישה לאנגלית.

הרשמה

תלמיד/ה שאינו/ה מאוניברסיטת תל אביב מתבקש/ת למלא את הטופס בקישור

https://docs.google.com/forms/d/e/1FAIpQLSeyex9Ijr6WaNJ_GYq9lvqv1UI05QSE8p7yCqwcraOQLtSxIA/viewform?usp=sf_link

אוניברסיטת תל אביב תממן את הוצאות הנסיעה של המרצים והתלמידים שאינם מאוניברסיטת תל אביב. יש למלא את הטופס המתאים (ניתן לקבלו מגב' נורית ליברמן), ולהגיש את הקבלות (מקור ולא צילום). ללא קבלות לא ניתן יהיה לקבל החזר הוצאות. את הטופס עם הקבלות יש למסור לגב' נורית ליברמן, במזכירות הפקולטה.

לפרטים נוספים ניתן לפנות למיכה מנדל micha.mandel@mail.huji.ac.il, יאיר גולדברג
yair.goldy@gmail.com או מלכה גורפיין malkago12@gmail.com.

נושאי הקורס: נתונים של סדרות זמן נפוצים בתחומים שונים, כולל פיננסים, כלכלה, הנדסה ומדעי החברה. במסגרת הקורס יוצגו מושגים בסיסיים, טכניקות מידול ושיטות חיזוי לניתוח סדרות זמן, הן מנקודת מבט קלאסית והן מנקודת מבט של למידת מכונה. הנושאים כוללים הבנת נתוני סדרות זמן, סטציונריות, מודלי Seasonal Autoregressive Integrated Moving Average, ניתוח במרחב התדר, קלמן פילטרס, טיפול בנתוני אורך, וחיזוי באמצעות כלים קלאסיים ולמידת מכונה.

Course Description:

This course provides an introduction to the analysis and forecasting of time series data. Time series data is prevalent in various fields, including finance, economics, engineering, and social sciences. The course covers fundamental concepts, modeling techniques, and prediction methods for time series analysis, both from classical and machine learning perspectives. Topics include understanding time series data, stationarity, modeling with SARIMA (Seasonal Autoregressive Integrated Moving Average) models, frequency domain analysis, Kalman filters, handling longitudinal data, and prediction using classical and machine learning tools.

Course Syllabus:

Module 1: Introduction to Time Series Analysis

- What is a time series?
- Types and characteristics of time series data
- Time series components: trend, seasonality, and noise
- Basic exploratory data analysis for time series

Module 2: Modeling Time Series with SARIMA Models

- Autoregressive (AR) models
- Moving Average (MA) models
- Autoregressive Moving Average (ARMA) models
- Seasonal ARIMA (SARIMA) models
- Model identification, estimation, and diagnostic checking

Module 3: Frequency Domain Analysis

- Fourier analysis and Fourier transform
- Periodogram and spectral analysis
- Power spectral density estimation
- Filtering in the frequency domain

Module 4: Kalman Filters for Time Series Analysis

- Introduction to Kalman filters
- State-space models and the Kalman filter equations
- Filtering and smoothing with the Kalman filter
- Applications in time series analysis and forecasting

Module 5: Longitudinal Data Analysis

- Understanding longitudinal data
- Handling panel data and repeated measures

Mixed-effects models for longitudinal data

Longitudinal forecasting techniques

Module 6: Prediction using Classical Methods

Exponential smoothing models (Simple, Holt's, and Winter's methods)

Autoregressive Integrated Moving Average (ARIMA) models

Forecast evaluation and accuracy measures

Model selection and model diagnostics

Module 7: Prediction using Machine Learning Tools

Introduction to machine learning for time series analysis

Regression-based methods (linear regression, support vector regression, etc.)

Neural networks for time series forecasting (e.g., LSTM, GRU)

Evaluation and comparison of machine learning models

Course Project:

Students will complete a time series analysis and forecasting project using real-world data. The project will involve data preprocessing, model selection, estimation, and evaluation of forecast accuracy.

Prerequisites:

Basic knowledge of statistics and probability

Familiarity with regression analysis and hypothesis testing

Proficiency in a programming language (e.g., Python, R) for data analysis

Assessment:

Assignments throughout the course

Course project

Note: The syllabus is subject to modification and can be tailored according to the specific needs and time constraints of the course.

תיאוריה של למידה

ד"ר עמיחי פיינסקי, אוניברסיטת תל אביב

הקורס יעסוק ביסודות התיאורטים של למידת מכונה מודרנית. ננתח את האתגרים הבסיסיים של בעיות למידה ונלמד כלים מגוונים להתמודד איתם. הקורס יכלול נושאים מתקדמים בסטטיסטיקה, הסתברות ואנליזה קמורה.

Topics:

Weeks 1-2: Introduction to learning theory.

Weeks 3-4: Finite hypothesis class.

Week 5: Rademacher complexity.

Weeks 6-7: Linear hypothesis classes.

Weeks 8: VC theory and Sauer's lemma.

Week 9: Growth functions and VC dimension.

Week 10: PAC-Bayes bounds.

Week 11: Algorithmic stability.

Reading:

Ross and Peko, "A Second Course in Probability".

Grimmett and Stirzaker, "Probability and Random Processes".

A tutorial on statistical learning theory by Bousquet, Boucheron, and Lugosi.

Bartlett and Mendelson, "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results", 2002.

McAllester, "Simplified PAC-Bayesian Margin Bounds", 2003.

דרישות הקורס: תרגילי בית, בחינה סופית.

הרכב הציון הסופי: 20% שעורי בית, 80% בחינה סופית.

מטרת הקורס היא להציג כמה מהאתגרים הסטטיסטיים העיקריים שעידן Big Data מביא, ולדון בפתרונות שלהם. הקורס הוא קורס "נושאים" שבו נחליף תחום עיסוק מדי מספר שבועות. נדון בנושאים השונים בצורה רחבה אך תוך ירידה ספציפית לצדדים הסטטיסטיים הטכניים שלהם. נביא בחשבון גם אספקטים יישומיים וחשובים. בהתאם לכך, שיעורי הבית ופרויקט הסיום יכללו שילוב בין עבודה פרקטית בתכנות ועם נתונים, לבין ניתוחים תיאורטיים.

The goal of this course is to present some of the unique statistical challenges that the new era of Big Data brings, and discuss their solutions. The course will be a topics course, meaning we will discuss various aspects that may not necessarily be related or linearly organized. Our challenge will be to cover a wide range of topics, while being specific and concrete in describing the statistical aspects of the problems and the proposed solutions, and discussing these solutions critically. We will also keep in mind other practical aspects like computation and demonstrate the ideas and results on real data when possible. Accordingly, the homework and the final project will include a combination of hands-on programming and modeling with theoretical analysis.

Big Data is a general and rather vague term, typically referring to data and problems that share some of the following characteristics:

- It is big (obviously), this could mean having many observations (large n), many features/variables (large p), or both.
- It has additional structure information: temporal, spatial, graph structure (like network data), etc.
- It leads to non-traditional modeling problems, like network evolution, collaborative filtering, structured learning, etc.
- It presents significant practical challenges in handling the data and modeling it, including:
 - The need to maintain privacy and security of the data while sharing it and extracting information from it.
 - The difficulty in storing and performing calculations at scale.
 - The difficulty in correctly interpreting the data and generating valid statistical modeling problems from it.
- The full extent of its potential utility is unclear and subject to research.

Some examples of typical Big Data domains gaining importance in recent years:

- Internet usage data, including social network information, search and advertising information, etc.
- Health records and related information.
- Scientific databases, including areas like particle physics, electron microscopy and genetics.
- Images and video surveillance data.

A key topic in data modeling in general and Big Data in particular is predictive modeling (regression, classification). Since the course Statistical Learning deals mainly with exposition and statistical analysis of algorithms in this area, it will not be a focus of this course. However, some aspects of this area that are not covered in that course, in particular the $p \gg n$ case, efficient computation, and deep learning, will be discussed in some detail.

Tentative list of topics to be covered during the semester:

- Network modeling: Probabilistic models of network evolution; Parameter estimation and inference.
- Privacy: Differential privacy; Algorithms to guarantee privacy in different settings; Examples of privacy breaches.
- Efficient computation in predictive modeling: Regularization path algorithms; Stochastic gradient descent.
- Statistical validity of scientific research on modern data: Replicability; Sequential testing on public databases
- Spectral analysis of large random matrices: statistical and computational issues.
- $p \gg n$: Sparsity and computation.
- Deep learning: theory and methodology.
- Turning data into modeling: Competitions and proof of concept projects; Leakage in data mining

We will have several guest lectures (as time permits) during the semester, but they will be treated as regular classes rather than enrichment classes (specifically, their material will be included in the homework and the final).