Short Courses of ISCA 2021 Applied Statistics Symposium

Statistical and algorithmic foundations of reinforcement learning

Yuting Wei, PhD, Assistant Professor in the Statistics and Data Science Department, Carnegie Mellon University

Yuejie Chi, PhD, Associate Professor in the Department of Electrical and Computer Engineering, Carnegie Mellon University

Yuxin Chen, PhD, Assistant Professor in the Department of Electrical and Computer Engineering, Princeton University

Zhengyuan Zhou, PhD, Assistant Professor at the Stern School of Business, New York University

Duration: One day

Abstract:

Reinforcement learning (RL) studies how an agent learns to make sequential decisions to improve its performance over time through its interactions with an unknown environment. In the past few years, rapid advances in algorithms and explosive growth of computational and memory resources have enabled tremendous empirical success. In this short course, we convey this well-founded excitement by presenting a modern tutorial on the foundation of RL, both as a field and as a set of ideas. Our tutorial underscores important statistical and algorithmic developments in RL, both new and classic.

We start by presenting a unified framework of the RL problem---from multi-arm bandits to general Markov decision processes---and then introduce classical planning algorithms when precise descriptions of the environments are available (Part I). Equipped with this background, we discuss two prominent approaches---model-based and model-free RL---assuming access to a generative model of the environment (Part II). Next, we discuss how to design intelligent exploration to optimize the agent's long-term performance when adaptively acting in a real environment (Part III). Finally, we present another distinctive approach: policy optimization, and conclude with further pointers to the literature (Part IV). Through this intense course, we hope to bring students and researchers in the statistical disciplines quickly up to speed to the heart of important insights that underlie the modern RL success.

About the instructors:

Dr. Yuting Wei is currently an Assistant Professor in the Statistics and Data Science Department at Carnegie Mellon University. Prior to that, she was a Stein Fellow at Stanford University, and she received her Ph.D. in statistics at University of California, Berkeley working

with Martin Wainwright and Aditya Guntuboyina. She was the recipient of the 2018 Erich L. Lehmann Citation from the Berkeley statistics department for her Ph.D. dissertation in theoretical statistics. Her research interests include high-dimensional and non-parametric statistics, statistical machine learning, and reinforcement learning.

Dr. Yuejie Chi is an Associate Professor in the department of Electrical and Computer Engineering, and a faculty affiliate with the Machine Learning department and CyLab at Carnegie Mellon University, where she held the Robert E. Doherty Early Career Development Professorship from 2018 to 2020. She received her Ph.D. and M.A. from Princeton University, and B. Eng. (Hon.) from Tsinghua University, all in Electrical Engineering. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning and inverse problems, with applications in sensing systems, broadly defined. Among others, Dr. Chi received the Presidential Early Career Award for Scientists and Engineers (PECASE), the inaugural IEEE Signal Processing Society Early Career Technical Achievement Award for contributions to high-dimensional structured signal processing, and was named a Goldsmith Lecturer by IEEE Information Theory Society.

Dr. Yuxin Chen is currently an assistant professor in the Department of Electrical and Computer Engineering at Princeton University, and is affiliated with Applied and Computational Mathematics, Computer Science, and Center for Statistics and Machine Learning. Prior to joining Princeton, he was a postdoctoral scholar in the Department of Statistics at Stanford University, and he completed his Ph.D. in Electrical Engineering at Stanford University. His research interests include high-dimensional statistics, mathematical optimization, and reinforcement learning. He has received the Princeton graduate mentoring award, the AFOSR Young Investigator Award, the ARO Young Investigator Award, the ICCM best paper award (gold medal), and was selected as a finalist for the Best Paper Prize for Young Researchers in Continuous Optimization.

Dr. Zhengyuan Zhou is an assistant professor at the Stern School of Business, New York University. Before joining NYU Stern, he spent the year 2019-2020 as a Goldstine research fellow at IBM research. He received his BA in Mathematics and BS in Electrical Engineering and Computer Sciences, both from UC Berkeley. Subsequently, he has received a Master's in Computer Science, a Master's in Statistics, a Master's in Economics and a PhD in Electrical Engineering (with minors in Mathematics and Management Science & Engineering), all from Stanford University in 2019. His research interests include contextual bandits, reinforcement learning and data-driven sequential decision making problems at large.

Instructors' Teaching Experience:

Dr. Yuting Wei teaches at the Statistics and Data Science Department at Carnegie Mellon University, where she has designed 2 graduate courses on high dimensional statistics and large-scale optimization and 1 undergraduate course on introduction to probability.

Dr. Yuejie Chi teaches regularly at Carnegie Mellon University, and has presented several invited tutorials at conferences such as ICASSP, ITW and EUSIPCO.

Dr. Yuxin Chen has developed 5 graduate courses and 1 undergraduate course at Princeton University. The courses he has taught has been included in the Princeton Engineering Commendation List for Outstanding Teaching for four times. He has co-taught three invited tutorials at leading conferences in information theory and signal processing.

Dr. Zhengyuan Zhou teaches regularly at Stern School of Business at New York University. He has developed 2 graduate courses (convex optimization and machine learning), 1 MBA course and 1 undergraduate course on decision models and analytics.

Biomarker discovery and pathway enrichment analysis of omics data

Ali Rahnavard, PhD, Assistant Professor of Biostatistics and Bioinformatics, George Washington University

Himel Mallick, PhD, Senior Scientist with Merck Research Laboratories, Merck & Co., Inc

Duration: Half-Day

Abstract:

Methodological advancements paired with measured multiomics data using high-throughput technologies enable capturing comprehensive snapshots of biological activities. In particular, low-cost, culture-independent omics profiling has made metagenomics, metabolomics, and proteomics ("multiomics") surveys of human health, other hosts, and the environment. The resulting data have stimulated the development of new statistical and computational approaches to analyze and integrate omics data, including human gene expression, microbial gene products, metabolites, and proteins, among others.

Multiomics data generated from diverse platforms are often fed into generic downstream analysis software without proper appreciation of the inherent data differences, resulting in incorrect interpretations. Further, there are also an extensive collection of downstream analysis software platforms, and appropriately selecting the best tool can be extraneous for untrained researchers.

This workshop will thus present a high-level introduction to computational multiomics, highlighting the state-of-the-art in the field and outstanding challenges geared towards downstream analysis methods. The workshop will include introducing typical multiomics studies' biological goals and the statistical methods currently available to achieve them.

Statistical methods for composite time-to-event outcomes

Lu Mao, PhD, Assistant Professor in the Department of Biostatistics and Medical Informatics (BMI), University of Wisconsin-Madison

Duration: Half-Day

Abstract:

This course provides an overview of the recent statistical methodology developed for the analysis of composite time-to-event outcomes. Such outcomes typically combine death and (possibly recurrent) nonfatal events, such as hospitalization, tumor progression, or infection, and are routinely used as the primary efficacy endpoint in modern phase-III clinical trials. The traditional approach to composite outcomes is to focus on time to the first event, whichever type it is. Recent years have seen a surge of novel statistical methods, attracting the attention of both methodologists and practitioners. These include the win ratio (Pocock et al., 2012) and its various extensions, the restricted mean time in favor of treatment (an extension of the restricted mean survival time), generalized semiparametric proportional odds regression models, and so on. The new methods improve upon the existing ones in:

1. Proper prioritization of death over nonfatal events;
2. Fuller utilization of multiple/recurrent events;
3. Clear and interpretable definition of estimands in compliance with the ICH E9 (R1) guideline;
4. Versatile modeling strategies for general outcome types

In addition, a number of user-friendly R-packages that implement the new methods have also become available. This course will supplement methodological studies with hands-on analysis of real-world data using R.

About the instructor:

Dr. Lu Mao joined the Department of Biostatistics and Medical Informatics (BMI) at UW-Madison as an Assistant Professor after finishing his doctoral dissertation on regressions of competing risks data at UNC Chapel Hill in 2016. His current research interests include survival analysis, causal inference, semiparametric theory, and clinical trials. He currently serves as the PI of an NIH R01 grant on statistical methodology for composite time-to-event outcomes in cardiovascular clinical trials and an NSF grant on causal inference in randomized trials with noncompliance. Besides methodological research, he also engages in collaborative studies in cardiology, radiology, cancer, and health behavioral interventions, where time-to-event and longitudinal data are routinely collected.

Instructor's Teaching Experience:

Dr. Lu Mao taught a graduate-level course on applied longitudinal analysis (using Fitzmaurice, Laird & Ware (2011) as textbook) in Spring 2018 and has been serving as the instructor for the

graduate-level survival analysis course since 2017. Outside campus, he taught a half-day short course on statistical methods for composite endpoints at the 2019 ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop held in Washington, DC.

Large-Scale Spatial Data Science

Marc Genton, PhD, Distinguished Professor of Statistics at the Spatio-Temporal Statistics and Data Science (STSDS) research group, King Abdullah University of Science and Technology

Huang Huang, PhD, Research Scientist at the Spatio-Temporal Statistics and Data Science (STSDS) research group, King Abdullah University of Science and Technology

Sameh Abdulah, PhD, Research Scientist at the Extreme Computing Research Center (ECRC), King Abdullah University of Science and Technology

Duration: Half-Day

Abstract:

Spatial data science is concerned with analyzing the spatial distributions, patterns, and relationships of data over a predefined geographical region. It relies on the dependence of observations where the primary assumption is that nearby spatial values are associated in a certain way. For decades, the size of most spatial datasets was modest enough to be handled by exact inference. Nowadays, with the explosive increase of data volumes, High-Performance Computing (HPC) has become a popular tool for many spatial applications to handle massive datasets. Big data processing becomes feasible with the availability of parallel processing hardware systems such as shared and distributed memory, multiprocessors, and GPU accelerators. In spatial statistics, parallel and distributed computing can alleviate the computational and memory restrictions in large-scale Gaussian random process inference. In this course, we will first briefly cover the motivation, history, and recent developments of statistical methods so that the participants can have a general overview of spatial statistics. Then, the cutting-edge HPC techniques and their application in solving large-scale spatial problems with the new software ExaGeoStat will be presented.

About the instructors:

Marc G. Genton is a Distinguished Professor of Statistics at the Spatio-Temporal Statistics and Data Science (STSDS) research group, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. He received his Ph.D. in Statistics in 1996 from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He also holds an M.S. degree in Applied Mathematics teaching from EPFL. Prior to joining KAUST, he held faculty positions at MIT, North Carolina State University, the University of Geneva, and Texas A&M University. Most of Prof. Genton's research centers around spatial and spatio-temporal statistics. His interests include the statistical analysis, visualization, modeling, prediction, and uncertainty quantification of spatio-temporal data, with applications in environmental and climate science, renewable energies, geophysics, and marine science. He is a Fellow of the American Statistical Association, of the Institute of Mathematical Statistics, of the American Association for the Advancement of Science, and an elected member of the International Statistical Institute. In 2010, he received the El-Shaarawi award for excellence from the International Environmetrics

Society (TIES) and the Distinguished Achievement Award from the Section on Statistics and the Environment (ENVR) of the American Statistical Association (ASA). In 2017, he received the Wilcoxon Award for Best Applications Paper in the journal Technometrics. He is the 2020 Georges Matheron Lecturer of the International Association for Mathematical Geosciences. Personal webpage: http://stsds.kaust.edu.sa

Huang Huang is a research scientist at the Spatio-Temporal Statistics and Data Science (STSDS) research group, King Abdullah University of Science and Technology (KAUST), Saudi Arabia. Before working at KAUST, Huang was a postdoctoral fellow at the National Center for Atmospheric Research (NCAR), the Statistical and Applied Mathematical Sciences Institute (SAMSI), and Duke University, focusing on spatio-temporal statistical inference for large climate data sets. He received his Ph.D. in Statistics in 2017 from KAUST and M.S. and B.S. in Mathematics in 2014 and 2011 from Fudan University, China. His research interests include spatio-temporal statistics, functional data analysis, Bayesian modeling, machine learning, and High-Performance Computing (HPC) for large climate data. Personal webpage: https://hhuang90.github.io

Sameh Abdulah is a research scientist at the Extreme Computing Research Center (ECRC), King Abdullah University of Science and Technology, Saudi Arabia. Sameh was a postdoctoral fellow at ECRC from 2016 to 2019. He received his M.S. and Ph.D. degrees from the Ohio State University, Columbus, USA, in 2014 and 2016. His work is centered around High-Performance Computing (HPC) applications, bitmap indexing in big data, large spatial datasets, parallel statistical applications, algorithm-based fault tolerance, and machine learning and data mining algorithms. Sameh is also a reviewer in several HPC-related journals and conferences. Personal webpage: https://sites.google.com/view/samehabdulah

Instructors' Teaching Experience:

Marc has 25 years of teaching experience at universities, including short-courses.

Huang has several short-course teaching experiences, especially for R tutorials and hands-on exercises.

For six years, from 2006 to 2012, Sameh taught several computer science-related courses to undergraduate students such as Computer Programming, Data Structures, Algorithms, Object-Oriented Programming, Theory of Computing, Artificial Intelligence, Data Science, Machine Learning, Software Engineering, Cloud Computing, Computer Security, and Big Data. His experience in teaching involves leading theoretical classes and practical exercises.

Functional Data Analysis Using R

Jiguo Cao, PhD, Canada Research Chair in Data Science and Professor at the Department of Statistics and Actuarial Science, Simon Fraser University

Duration: One-Day

Abstract:

Functional data analysis (FDA) is a growing statistical field for analyzing longitudinal trajectories, curves, images or any manifold objects. FDA treats each random function observed over the whole domain as a sample element. Functional data can be commonly be found in many applications such as fitness data from the wearable device, air pollution, longitudinal studies, time-course gene expressions and brain images. This short course will cover the major FDA methods such as functional principal component analysis and functional linear regression models. All these methods will be demonstrated with real data applications using the R programming language. The goal of this short course is that trainees can learn how to use and develop FDA methods to analyze data.

About the instructor:

Dr. Jiguo Cao is the Canada Research Chair in Data Science and Professor at the Department of Statistics and Actuarial Science, Simon Fraser University. He is a prolific researcher who addresses critical, data-driven applications with well-tuned new models and estimation methods. He is a dedicated scholar with a worldwide reputation, a collaborator with a rich array of scientists, and an outstanding mentor of future scholars. Over a short period of time, Dr. Cao has made inroads across an impressively diverse range of sub-disciplines of statistics and data science; he builds on the synergies between functional data analysis (FDA), differential equations and big data to unveil novel statistical methods that greatly enhance users' understanding of their data. His high-impact, statistical frameworks are applied to real-world problems across various disciplines, including neuroscience, pharmacology, genetics, ecology, environment, and engineering. Dr. Cao's research is highly visible, with over 70 refereed publications and the delivery of 9 prestigious short courses on FDA and 74 invited talks and seminars in Australia, Canada, China, and USA.

Instructor's Teaching Experience:

Dr. Jiguo Cao has delivered 9 prestigious short courses on FDA at international conferences, universities and companies in Canada, China, and USA.

Mining Electronic Health Record (EHR) data: the past, presence, and future

Shuangge Ma, PhD, Professor of Biostatistics, Yale School of Public Health

Duration: Half-Day

Abstract:

With the fast development of information technologies and computing power, many nations (both central and local governments), large insurance systems, and health care systems have established or are establishing large electronic health record (EHR) databases. Effectively mining such databases using advanced data science techniques can generate unprecedented value for public health care, biomedicine, insurance management and planning, and other purposes. In this course, we will review the past and current status of the establishment of EHR databases and discuss successful (and unsuccessful) examples. Simple and advanced data science techniques that have been developed tailored to EHR data will be reviewed, with extensive examples provided. Discussions will then be provided on the future of EHR data mining, what needs to be done in terms of database/system development and data science methodology development, their implications, and the unique role statisticians can play.

About the instructors:

Shuangge Ma is Professor of Biostatistics at Yale University. His research interests include big data, EHR (electronic health record) data analysis, network analysis, public health, and health economics. He is an Elected Member of ISI and Fellow of ASA. He has published over 200 articles in prestigious journals and 2 books. He has been playing a leading role in biostatistical studies on cancer, health economics, and other areas. He has taught multiple short courses on advanced data science technologies at international conferences and renowned universities. He is an associate editor of JASA and other seven prestigious journals.