Causal Feature Selection
and its Applications in Biomedical and Health Data Science

Sisi Ma, Ph.D., Assistant Professor
Department of Medicine and Institute for Health Informatics
University of Minnesota, Minneapolis, MN
Constantin F. Aliferis, M.D., Ph.D. , Professor and Director
Institute for Health Informatics at University of Minnesota, Minneapolis, MN
Alexander Statnikov, Ph.D. Adjunct Associate Professor
Dept of Medicine at New York University Langone Medical Center, NY, NY

Workshop type: instructional

Workshop length: half-day (3 hours)

Worksop Faculties:

Sisi Ma, Ph.D. Assistant Professor in Department of Medicine and Institute for Health Informatics at University of Minnesota, Minneapolis, MN

Alexander Statnikov, Ph.D. is an Adjunct Associate Professor in the Department of Medicine at New York University Langone Medical Center, NY, NY

Constantin F. Aliferis, M.D., Ph.D. Professor and Director of the Institute for Health Informatics at University of Minnesota, Minneapolis, MN

Workshop Content

Predictive modeling has become an important tool in biomedical and health science over the past few years. Applications of predictive modeling in biomedical and health science span a wide range. However, developing high quality, generalizable, and easily deployed models for biomedical and health science applications remains challenging especially when using Big Data. The workshop will address how causal feature selection methods tackle the feature selection problem, a critical component of predictive modeling.

Causal feature selection methods, also known as Markov Boundary or Markov Blanket (MB for short) methods are specifically suitable for data from the health science domain where large numbers of features are measured and highly parsimonious and maximally predictive models are sought. Causal feature selection methods methods are particularly robust to situations with relatively small number of samples (e.g. multi-omics data for cancer patients, clinical trial data, data from mental health domain). Causal feature selection methods have the distinct advantage that they identify features that, under broad conditions, have causal and predictive rather than just predictive interpretations compared to features selected by non-causal feature selection methods. Furthermore, causal features are more robust to distribution shifts and can therefore, under conditions, generate accurate prediction on data collected from populations that are different from the training population.

In this workshop, we will introduce: (1) predictive modeling and the key challenges for predictive modeling in Big Data health data science, (2) Classic and state of the art non-causal feature selection methods (3) the theoretical background for causal discovery from observational data, and for optimal feature selection which is the foundation for causal feature selection methods, (4) causal feature selection algorithms for both binary and survival outcomes (5) hands-on and interactive tutorial for executing causal feature selection methods in R. The benefit of causal feature selection methods in comparison to non-causal feature selection methods will be illustrated with examples using analysis on real data from different domains of biomedicine.

Workshop Outline:

The duration of the tutorial will be 3 hrs, with 1 hr and 20 min devoted to lectures, 1 hr and 20 min devoted to hands-on tutorial and Q&A session, and 20 min break in between.

Part I Predictive Modeling : Special Challenges in Health Sciences (30 min)

- Brief introduction to predictive modeling/supervised learning including binary prediction models and survival models.
- State of the art non-causal feature selection methods: e.g. Lasso, recursive feature elimination
- Challenges of feature selection for data from the health science domain and properties of non-causal feature selection methods.

Part II Causal Feature Selection Methods: (50 min)

- Introduction to theory of causal structural learning and causal feature selection
- Introduction to Optimal feature selection theory
- Conditional Independence Tests for Different Distributions
- Causal feature/Markov Boundary selection algorithms
- Properties and advantages of causal feature selection methods

Part III Hands-on Tutorial for causal feature selection: (1 hr and 20 min)

- Case study with using high-throughput genomics data to predict cancer diagnostics (35 min; R)
    - Comparison of predictive performance for cancer diagnostics between models built with non-causal and causal feature selection methods
    - Comparison of number of features selected for cancer diagnostics between models built with non-causal and causal feature selection methods
    - Compare features selected by causal features selection vs non-causal feature selection method, discuss the differences with respect to the information content of the features.
- Case study with a psychiatry dataset for predicting PTSD (10 min; R)
    - Comparison of predictive performance for PTSD diagnostics between models built with non-causal and causal feature selection methods
    - Comparison of number of features selected for PTSD diagnostics between models built with non-causal and causal feature selection methods
    - Compare features selected by causal features selection vs non-causal feature selection method, discuss the differences with respect to the information content of the features.
- Case study with using gene expression and clinical data to predict survival time (35 min; R)
    - Comparison of predictive performance for time-to-death between models built with non-causal and causal feature selection methods
    - Comparison of number of features selected for time-to-death between models built with non-causal and causal feature selection methods
    - Compare features selected by causal features selection vs non-causal feature selection method, discuss the differences with respect to the information content of the features.

Description of the Interactive Component:

The interactive component of the workshop constitutes of a hands-on tutorial for causal feature selection and a Q&A session. For the hands-on tutorial, we will use R to execute all analyses. The R scripts/notebooks and datasets used in the workshop will be distributed to the registered participants prior to the workshop. A list of R packages required to execute the analysis will be supplied with instructions for installing the R packages. We highly recommend the participants to install these packages prior to the workshop if they are interested in running the analysis during the

workshop. For the Q&A session, workshop faculties will answer participants' questions regarding causal feature selection.

Specific educational objectives or outcomes:

The first part of the workshop will introduce relevant theory and methods regarding predictive modeling and feature selection with their applications in biomedical and health science, especially Big Data. Big data from biomedical and health domain is complex (variables from different and complex distributions and forms complex joint distributions), high dimensional (reaching millions of variables), and the variables often contain overlapping information. The characteristics of the data poses significant challenges for feature selection, i.e. identifying which features to include in the predictive model. Non-causal feature selection algorithms may select features that generate decent predictive models in the training population (e.g. patients in one particular hospital) but the model is likely to produce unsatisfactory performance when applied to a different population (e.g. patients in another hospital) due to distribution shift.

Also, non-causal feature selection methods select features based on their predictive relevance which may result in selecting multiple variables that contain redundant information, which results in costly model deployment. These specific challenges for predictive modeling and feature selection for biomedical data will be discussed in detail and illustrated with examples.

The second part of the workshop will aim to resolve some of the challenges for predictive modeling described in the first part of the tutorial by introducing causal feature selection techniques. Causal feature selection methods select the variables that are causally relevant to the target of interest. As long as the causal relationship among the selected variables and the target variable remains unchanged, the predictive model can be easily generalized to generate predictions in different populations. Further, the causal model selection methods will not select variables that are correlated with the target but not (directly) causally relevant, resulting in a smaller selected variable set compared to the non-causal methods. A smaller variable set makes the deployment of the causal model more cost-efficient. We will highlight the advantages of causal models and introduce several causal feature selection methods.

The third part of the workshop is designed to give the audience hands-on experience of executing some of the causal feature selection methods on datasets from the biomedical and health science domain. In this part, the audience will follow a data analytic pipeline and protocol for predictive modeling with causal and non-causal feature selection methods. Tips and pointers will be given regarding model diagnostics and model interpretation as well as different software implementation of the algorithms.

The knowledge gained in this workshop will allow attendees to gain an in-depth understanding of predictive modeling, causal modeling, and causal feature selection. Researchers would be able to conduct predictive modeling on complex and high-dimensional data with various feature selection methods and learn benchmarking, model diagnostic, and model interpretation techniques. The specific learning objectives are: (1) Understand the advantages and limitations of non-causal and causal feature selection techniques. (2) Gain hands-on experience with predictive modeling and causal feature selection methods by following the principles and protocols presented.

Intended Audiences:

Biomedical and clinical informatics scientists and researchers, clinicians, biostatisticians, data analysts, data scientists of various other backgrounds, and scientific programmers. We anticipate that the workshop content will be 25% basic, 50% intermediate, and 25% advanced for the expected attendees.

Prerequisites Knowledge:

Prior knowledge of binary classification and survival modeling is not required but will facilitate deeper understanding of materials presented in the workshop. Basic skills in R programming would maximize the benefit from the hands-on portion of the tutorial.

Experience of the workshop instructors/leaders in in the targeted content areas:

Sisi Ma, Ph.D. (Instructor,Organizer) is an Assistant Professor in the Department of Medicine and Institute for Health Informatics at University of Minnesota. Dr. Ma's research interest is machine learning and causal modeling with special focus on their applications in biological and medical data. Dr. Ma has extensive experience in all stages of predictive modeling (feasibility analysis, model design, feature construction and selection, prototyping, implementation, model diagnostics, validation, performance evaluation, model interpretation and model deployment) for datasets from the health domain. Dr. Ma has also designed, implemented, and tailored multiple computational

causal discovery and causal feature selection algorithms for health science data. Dr. Ma has published more than 40 papers in peer reviewed journals and conferences proceedings and presented her work in multiple national and international conferences to diverse audiences. Dr. Ma has taught graduate courses in machine learning and computational causal discovery. She served on the organizing committee of the 4th Workshop on Data Mining for Medical Informatics, Causal Inference for Health Data Analytics. She has also organized a causal feature selection tutorial in Medinfo 2017. The current tutorial builds on the Medinfo tutorial and included more analytical experiments on real datasets for not only binary but also survival outcome prediction.

Constantin F. Aliferis, M.D., Ph.D., FACMI (Organizer) Dr. Aliferis holds the Positions of Professor of Medicine and Data Science and is also the Director of the Institute for Health Informatics, and the first University of Minnesota Chief Research Informatics Officer. Dr. Aliferis is also Chief Data Analytics Officer for MHealth (collectively the Academic Health Center, Fairview Health Systems, and University of Minnesota Physicians). Dr. Aliferis' research is focused on inventing and applying high dimensional modeling and analysis algorithms, protocols and systems designed to transform biomedical data into novel actionable scientific knowledge. His three key areas of broad interest are (a) use of advanced informatics and analytics to accelerate and enhance the sophistication, volume, quality, and reproducibility of scientific research; (b) quality and cost improvements of healthcare using Big Data approaches; and (c) precision medicine. Dr. Aliferis has served previously as faculty and held leadership roles at Vanderbilt University and NYU where among other roles, he served as Founding Director of the NYU Center for Health Informatics and Bioinformatics, and held professorships in Biomedical Informatics, Computer Science, Cancer Biology, Biostatistics, Computational Biology, Pathology and Data Science.

Alexander Statnikov, Ph.D. (Organizer) is an Adjunct Associate Professor in the Department of Medicine and Center for Health Informatics and Bioinformatics at New York University Langone Medical Center and the leader for Machine Learning at American Express. He was previously the Director of the Computational Causal Discovery Laboratory, and the Benchmarking Director of the Best Practices Integrative Informatics Consultation Service at NYULMC. Dr. Statnikov designed many innovative causal feature selection algorithms, conducted their empirical evaluations, and made other important contributions to the fields of machine learning and pattern recognition, analysis of high-throughput biomedical data, computational causal discovery, and biomedical informatics.