JOINT PROGRAM IN SURVEY METHODOLOGY

Introduction to Big Data and Machine Learning for Survey Researchers

Monday, May 17, 2021 - Friday, May 28, 2021

Workshop for JPSM, University of Maryland

TRENT BUSKIRK, Ph.D.

Novak Family Distinguished Professor of Data Science
Bowling Green State University, Bowling Green, OH

COURSE OBJECTIVES

This course will provide survey and social science researchers with a broad overview of big data and opportunities it can provide for study design and analysis. We will also spend considerable time focusing on how to apply machine learning methods to analyze and visualize such data and focus our attention specifically on how machine learning methods can be applied to various aspects of survey design and analysis. This course will provide participants:

- an overview of key Big Data terminology and concepts
- an introduction to common data generating processes
- a discussion of some primary issues with linking Big Data with Survey Data
- brief discussion of coverage and measurement errors within the Big Data context
- a discussion of information extraction and signal detection in the context of Big Data
- a discussion of the similarities and differences in model building for inference versus prediction
- an overview of general concepts from machine learning as they apply to processing Big Data
- a discussion of signal detection and information extraction
- a discussion of the potential pitfalls for inference from Big Data
- an introduction to common machine learning methods and how to implement these in R including:
    - K-means and Hierarchical Clustering
    - K-nearest neighbors
    - Classification and Regression Trees
    - Random Forests
- A detailed set of examples discussing how these methods and approaches can be used within the entire survey research process from sample design to questionnaire design to survey weighting and to survey analysis.

WHO SHOULD ATTEND

Individuals in government, business, academia, and non-profit organizations who are interested

in big data applications to survey and social science research. Individuals who are interested in taking their understanding of machine learning methods from no or basic understanding to a level where the methods are applied in R are also encouraged to attend. This course provides a condensed overview of big data and the more common

machine learning methods. **Students are expected to have a very basic familiarity with the statistical software R (e.g. how to load a package and how to launch it in R).**

SUGGESTED READING

James, G., Witten, D., Hastie, T. and Tibshirani, R. *(2013) An Introduction to Statistical Learning with Applications in R*. Free PDF Version here: http://bit.ly/1iUJso0

Shmueli, G. (2010) "To explain or to predict?" *Statistical Science,* Vol. 25, No. 3, 289–310; https://arxiv.org/pdf/1101.0891.pdf

Buskirk, T.D. Kirchner, A., Eck, A. Signorino, C. (2018) "An Introduction to Machine Learning Methods for Survey Researchers" *Survey Practice,* Vol. 11(1). https://www.surveypractice.org/article/2718-an-introduction-to-machine-learning-methods-for-survey-researchers

THE INSTRUCTOR

TRENT D. BUSKIRK is the Novak Family Professor of Data Science and Chair of the Applied Statistics and Operations Research Department in the College of Business at Bowling Green State University. He is also a Fellow of the American Statistical Association and most recently served as the Conference Chair for the American Association of Public Opinion Research. Trent was also on the Scientific Committee for the BigSurv18 conference that was held this past year in Barcelona. Trent's research interests include applications of machine learning to the design and analysis of both probability and nonprobability based surveys, use of auxiliary data to improve address and telephone based sample survey designs and the use of technology to improve and enhance data collection. Trent's work has appeared in numerous journals including the Journal of Survey Statistics and Methodology, Public Opinion Quarterly, Social Science Computer Review, Field Methods and the Journal of Official Statistics among others. When he is not conducting research or teaching you can find Trent playing a very competitive game of pickleball!

COURSE MATERIALS

Registrants will be provided with a course lecture notebook.

JOINT PROGRAM IN SURVEY METHODOLOGY

COURSE SCHEDULE:

We will have a total of 14 videos to be viewed for the workshop over a two-week period beginning Monday May 17 through Friday May 21(Week 1) and Monday May 24through Friday May 28 (Week 2).

We will have two one-hour live Q and A session on Friday May 21 and Friday May 28 from Noon to 1 pm Eastern.

Video Segment (Week)	Topic
Video 1 – Week 1	Introduction to Big Data
Video 2 - Week 1	Introduction to Big Data, Part 2
Video 3 – Week 1	Working With Big Data – Properties, Promise and Peril
Video 4 – Week 1	Opportunities for Big Data in Survey Research
Video 5 – Week 1	Introduction to Machine Learning: Comparing Classical Statistics to Machine Learning Models – Part 1
Video 6 – Week 1	Introduction to Machine Learning: Comparing Classical Statistics to Machine Learning Models – Part 2
Video 7 – Week 1	Introduction to Machine Learning: Comparing Classical Statistics to Machine Learning Models – Plotting Big Data and Model Validation 1
Video 8 – Week 1	Introduction to Machine Learning: Comparing Classical Statistics to Machine Learning Models – Model Validation Part 2
Video 9 – Week 1	Introduction to Machine Learning: Comparing Classical Statistics to Machine Learning Models – Model Evaluation Metrics
Video 10 – Week 2	Popular Machine Learning Algorithms: Intro and K-Means Clustering
Video 11 – Week 2	Popular Machine Learning Algorithms: Hierarchical Clustering
Video 12 – Week 2	Popular Machine Learning Algorithms: K-nearest Neighbors
Video 13 – Week 2	Popular Machine Learning Algorithms: CARTS
Video 14 – Week 2	Popular Machine Learning Algorithms: Random Forests