Syllabus

Introduction to Machine Learning and Big Data (ML I)
1 credits/2 ECTS

Prof. Trent D. Buskirk, Prof. Frauke Kreuter
Video lecture by Trent D. Buskirk, Frauke Kreuter

March 1st – March 22nd, 2021

Short Course Description

The amount of data generated as a by-product in society is growing fast including data from satellites, sensors, transactions, social media and smartphones, just to name a few. Such data are often referred to as "big data", and can be used to create value in different areas such as health and crime prevention, commerce and fraud detection. Big Data are often used for prediction and classification tasks. Both of which can be tackled with machine learning techniques. In this course we explore how Big Data concepts, processes and methods can be used within the context of Survey Research. Throughout this course we will illustrate key concepts using specific survey research examples including tailored survey designs and nonresponse adjustments and evaluation.

Course Objectives

This course covers…

- 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
- issues of coverage and measurement errors within the Big Data context
- inference versus prediction
- general concepts from machine learning including signal detection and information extraction
- potential pitfalls for inference from Big Data
- key analytic techniques (e.g. classification trees, random forests, conditional forests) to process Big Data using R with example code provided

Prerequisites

No prerequisites.

We recommend good understanding of the material typically taught in undergraduate statistics courses and some familiarity with regression techniques. Knowledge about

survey data collection at the level provided in the IPSDS course Fundamentals of Survey and Data Science.
While not a prerequisite, familiarity with the R software package (base R or R using Rstudio) is strongly encouraged.

Class Structure and Course Concept

This is an online course, using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video-recorded lectures and reading the required literature for each unit prior to participating in mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.
Just like in an on-site course, homework will be assigned and graded.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching videos, participating in online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb, you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. This is a 1-credit/2-ECTS course that runs for 4 weeks. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Monday, 1pm ET/7pm CET, starting March 1st, 2021

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on mannheim.instructure.com. If video participation via Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Sunday, 1pm ET/7pm CET).

Students have the opportunity to use a different Zoom meeting room to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Detailed information is posted on the course page in Canvas. Students are encouraged to post the times that they will be using the room to the course website forum to avoid scheduling conflicts. Students are not required to use Zoom and can use other online meeting platforms, such as Microsoft Teams, Google Hangout or Skype.

Grading

Grading will be based on:

- 4 online quizzes (worth 5% each)
- Participation in discussion during the weekly online meetings and submission of questions to the weekly discussion forums (deadline: Sunday, 1:00 PM EST/7:00 PM CET before class) demonstrating understanding of the required readings and video lectures (20% of grade). Obviously in the first week one question will be enough, since we just started.
- 3 homework assignments (worth 20% each)

A+ 100 - 97
A 96 - 93
A- 92 - 90
B+ 89 - 87
B 86 - 83
B- 82 - 80
Etc.

The grading scale is a base scale recommended by the MDM. Variations for grading on a scale are at the discretion of the instructor.

The final grade will be communicated under the assignment "Final Grade" in the Canvas course. Please note that the letter grade written in parentheses in Canvas is the correct final grade. The point-grade displayed alongside the letter grade is irrelevant and can be ignored.

Dates of when assignments will be due are indicated in the syllabus. Late assignments will not be accepted without prior arrangement with the instructors.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between the students and the instructors during the weekly online meetings. Therefore, we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description

The amount of data generated as a by-product in society is growing fast including data from satellites, sensors, transactions, social media and smartphones, just to name a few.

Such data are often referred to as "big data", and can be used to create value in different areas such as health and crime prevention, commerce and fraud detection. Big Data are often used for prediction and classification tasks. Both of which can be tackled with machine learning techniques. In this course we explore how Big Data concepts, processes and methods can be used within the context of Survey Research. Throughout this course we will illustrate key concepts using specific survey research examples including tailored survey designs and nonresponse adjustments and evaluation.

We will start with a discussion of key Big Data terminology and concepts. We place emphasis on understanding data generating processes and errors that can occur during these processes. Parallels between the errors in survey data collection and Big Data gatherings will be discussed. Special emphasis will be given to coverage error and measurement error. The key goal of any analytics task is information extraction and signal detection. Such task can look quite differently in the context of Big Data. We will compare common statistical methods to those use in the Big Data context and explain the difference in focus on prediction vs. causation. Most of the course time will be spend on general machine learning concepts, potential pitfalls, and the actual analytic processing steps when conducting applying techniques such as classification trees, random forests, conditional forests to process Big Data.

We use R and provide example code for the homework problems.

Readings

Primary Readings

There is no required textbook.

Useful recommended resources and readings

Data Mining Algorithms Explained Using R (2015) http://bit.ly/1yZYHjK

Online Resources for FREE lecture videos and labs in R http://bit.ly/1snBMk5

An overview of Machine Learning Functions available in R http://cran.r-project.org/web/views/MachineLearning.html

Recommended books on Machine Learning Methods and Related Topics

Data Mining for the Social Sciences (2015) http://bit.ly/1DpPFC2

An Introduction to Statistical Learning with Applications in R (2013)

Ratner, B. (2012) Statistical and Machine-Learning Data Mining, 2nd Ed. CRC Press, Boca Raton.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees. Pacific Grove: Wadsworth.

Lists of required and recommended readings for each class are provided below for each specific unit.

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland)

and in the MBS Honor Code, signed at the beginning of the program.

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Accessibility & Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

Students at the University of Mannheim should contact the Commissioner and Counsellor for Disabled Students and Students with Chronic Illnesses at http://www.uni-mannheim.de/studienbueros/english/counselling/disabled_persons_and_persons_with_chronic_illnesses/

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course (in addition to the standard university evaluation survey). Participation is entirely voluntary and highly appreciated.

Sessions

Week 1: Overview of Big Data; Working with Big Data; Classical Statistical Approaches versus Statistical Machine Learning

Video lecture: available Monday, February 22nd, 2021

Online meeting (Frauke Kreuter, Trent D. Buskirk): Monday, March 1st, 2021, 1pm ET/7pm CET

Online Quiz 1: due Wednesday, March 3rd, 2021, 1pm ET/7pm CET

Required Readings:
AAPOR (2015). AAPOR Report on Big Data.

Buskirk, T.D., Kirchner, A., Eck, A. and Signorino, C. (2018). An Introduction to Machine Learning Methods for Survey Researchers, *Survey Practice*, Vol. 11(1).

Kreuter, F., Peng, R. (2014). Extracting Information from Big Data: Issues of Measurement, Inference and Linkage. In Lane J. et al. (eds.) Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press. Manuscript

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science* 25 (3): 289–310.

Week 2: Model Evaluation/Validation; K-Means Clustering

Video lecture: available Monday, March 1st, 2021

Online meeting: Monday, March 8th, 2021, 1pm ET/7pm CET

Online Quiz 2: due Wednesday, March 10th, 2021, 1pm ET/7pm CET

Homework Assignment 1: Monday, March 15th, 2021, 1pm ET/7pm CET

Required Readings:
Molinaro, A. M., Simon, R., Pfeiffer, R.M. (2005). Prediction error estimation: a comparison of resampling methods. In *Bioinformatics*. 21(15):3301-7.

Ghani, R., Schierholz, M. (2017). Machine learning. In: I. Foster et al. (eds.). Big data and social science. A practical guide to methods and tools, Boca Raton: CRC Press, pp. 147-186. https://coleridge-initiative.github.io/big-data-and-social-science

Recommended Readings:

Schouten, B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39(1), 29-58.

Wagner, J. and Hubbard, F. (2014). Producing Unbiased Estimates of Propensity Models During Data Collection. *Journal of Survey Statistics and Methodology*, 2(3), 323-342.

Elliott, M. R. (2011). A Simple Method to Generate Equal-Sized Homogenous Strata or Clusters for Population-Based Sampling. *Annals of Epidemiology*, 21(4), 290–296. doi:10.1016/j.annepidem.2010.11.016

Krantz, A., Korn, R. and Menninger, M. (2009). Rethinking Museum Visitors: Using K-means Cluster Analysis to Explore a Museum's Audience. *Curator*, Vol. 52 (4), 363-374.

Barcaroli, G. (2014). Optimization of sampling strata with the SamplingStrata package.

Week 3: K Nearest Neighbors; CARTS

Video lecture: available Monday, March 8th, 2021

Online meeting: Monday, March 15th, 2021, 1pm ET/7pm CET

Online Quiz 3: due Wednesday, March 17th, 2021, 1pm ET/7pm CET

Homework Assignment 2: Monday, March 22nd, 2021, 1pm ET/7pm CET

Required Readings:

Buskirk, T.D. (2018). Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research. *Survey Practice*, Vol. 11(1).

Earp, M, Mitchell, M., McCarthy, J. and Kreuter, F. (2014). *Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey, Journal of Official Statistics*, Vol. 30(4), 701–719.

Recommended Readings:

Earp, M., Toth, D., Phipps,P. and Oslund, C. (2018). Assessing Nonresponse in a Longitudinal Establishment Survey Using Regression Trees. *Journal of Official Statistics*, Vol. 34, No. 2, 2018, pp. 463–481.

Phipps, Polly; Toth, Daniell. Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. Ann. Appl. Stat. 6 (2012), no. 2, 772--794. doi:10.1214/11-AOAS521.

Toth, D. 2017. rpms: Recursive Partitioning for Modeling Survey Data. R package version 0.2.0.

Week 4: Random Forests

Video lecture: available Monday, March 15th, 2021

Online meeting: Monday, March 22nd, 2021, 1pm ET/7pm CET

Online Quiz 4: due Wednesday, March 24th, 2021, 1pm ET/7pm CET

Homework Assignment 3: due Monday, March 29th, 2021, 1pm ET/7pm CET

Required Readings:
Buskirk, T.D. (2018). Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research. *Survey Practice*, Vol. 11(1).

Buskirk, T. D. & Kolenikov S. (2015). Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach*.

Recommended Readings:
Mendez, G., Buskirk, T.D., Lohr, S. and Haag, S. (2008). Factors Associated with Persistence in Science and Engineering Majors: An Exploratory Study Using Random Forests. Journal of Engineering Education, Vol. 97, No.1, pp. 57-70.