An Introduction to Big Data and Machine Learning for Survey Researchers and Social Scientists

Summer Institute in Survey Research Techniques, 2019
http://si.isr.umich.edu/

Instructor: Dr. Trent D. Buskirk
Time: July 10-July 12, 2019, 1:00 to 5:00 pm
Location: Mason Hall G449
E-mail buskirk@bgsu.edu
UM Office: 4034 ISR and Phone: 734-647-5380

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. An emerging practice in many areas is to append or link big data sources with more specific and smaller scale sources that often contain much more limited information. This practice has been used for some time by survey researchers in constructing frames by appending auxiliary information that is often not directly available on the frame, but can be obtained from an external source. Using Big Data has the potential to go beyond the sampling phase for survey researchers and in fact has the potential to influence the social sciences in general. Big Data is of interest for public opinion researchers and agencies that produce statistics to find alternative data sources either to reduce costs, to improve estimates or to produce estimates in a more timely fashion. However, Big Data pose several interesting and new challenges to survey researchers and social scientists among others who want to extract information from data. As Robert Groves (2012) pointedly commented, the era is “appropriately called Big Data and not Big Information”, because there is a lot of work for analysts before information can be gained from “auxiliary traces of some process that is going on in society.”

This course offers participants a broad overview of big data sources, opportunities and examples motivated within the survey and social science contexts including the use of social media data, para data and other such sources. This course also offers a detailed, practical introduction to four common machine learning methods that can be applied to big and small data alike at various aspects of a study’s lifecycle from design to nonresponse adjustments to propensity score matching to weighting and evaluation and analysis. The machine learning methods will be demonstrated in R and we will provide several different examples of using these methods along with multiple packages in R that offer these methods.

Evaluation
Grading for the course will be based on one practical project that will be due one week after the class ends. The project should be submitted individually in typewritten form and submitted electronically to buskirk@bgsu.edu with UMICH SUMMER 2019 in the subject line. The project will consist of two to three machine learning tasks to reinforce the examples we illustrate in class and to provide more hands on exposure to various machine learning methods we will cover in the course.
**Prerequisites**

Basic knowledge of descriptive statistics including an understanding of regression for both continuous and binary outcomes (e.g. linear and logistic regression) and a basic level of general exposure and basic experience with base R. Please bring a laptop computer to class with R or Rstudio installed. Please also load the following libraries which will be helpful throughout the class: rpart, rpart.plot, caret, randomForest, party, ggplot2, hclust, pROC, rattle and Rcmdr. You can install all of these packages at once by using the following line of code inserted at the R prompt:

```
install.packages(c('randomForest', 'party', 'rpart', 'rattle', 'Rcmdr','ggplot2', 'hclust','pROC', 'rattle','rpart.plot', 'caret'), dependencies=TRUE)
```

**Course Website**

The course website will be on CANVAS which can be accessed through [https://ctools.umich.edu/gateway/](https://ctools.umich.edu/gateway/). (In upper right hand corner – select CANVAS tab). The website contains assigned articles and pdfs of the course slides. Once the website is “published/open”...

- Students using a UMich e-mail can access the site instantly with their UMich e-mail address and password.

- Visiting students will be given a UMich username and password upon arrival. This is an important account as all details about the Summer Institute will come through this email.

- It is also possible for Visiting students and UMich students who wish to use a non-UMich e-mail (in addition to their UMich one) to access the course website by getting a Friend Account. A UMich Friend Account, a special kind of computer account that is used to give non-University of Michigan members access to the general University of Michigan web environment. You can use any e-mail address you want for your Friend Account, but this same e-mail address has to be registered with the class. Please see instructor for more information about this option.

**Course Outline**

**Day 1: Big Data: Potentials, Perils and Paradigms**

- An overview of Big Data, it’s potential, perils and total error framework.
- An overview of Machine Learning
- Understanding the differences between prediction and inference
- Challenges and some solutions for Plotting Big Data

**Day 2: Unsupervised Learning Methods**

- Challenges and Solutions for Plotting Big Data, Cont.
- K-Means Clustering
- Hierarchical Clustering
- k-nearest neighbors
- Hands on Exercises
Day 3: Supervised Learning Methods

- Regression and Classification Trees (unconditional and conditional)
- Ensemble Methods/Techniques
- Extra Trees
- Random Forests
- Hands on Exercises

Required Readings

  


Suggested Readings

Big Data, Privacy, Ethics and Mechanics

  

  

  
  [http://wpressutexas.net/cs378h/images/b/b3/LaneEtAlPrivacyBigDataAndThePublicGood.pdf](http://wpressutexas.net/cs378h/images/b/b3/LaneEtAlPrivacyBigDataAndThePublicGood.pdf)

Unsupervised Learning Applications

  
  doi:10.1016/j.annepidem.2010.11.016


Supervised Learning Applications