ME114 Introduction to Data Science and Big Data Analytics

http://github.com/kbenoit/ME114
LSE Methods Summer Programme 2015

Kenneth Benoit
Department of Methodology, LSE
kbenoit@lse.ac.uk

Slava Mikhaylov
University College London
s.mikhaylov@ucl.ac.uk

Version: August 17, 2015

Overview

Data Science and Big Data Analytics are exciting new areas that combine scientific inquiry, statistical knowledge, substantive expertise, and computer programming. One of the main challenges for businesses and policy makers when using big data is to find people with the appropriate skills. Good data science requires experts that combine substantive knowledge with data analytical skills, which makes it a prime area for social scientists with an interest in quantitative methods. This course integrates prior training in quantitative methods (statistics) and coding with substantive expertise and introduces the fundamental concepts and techniques of Data Science and Big Data Analytics.

Typical students will be Masters and PhD students from any field requiring the fundamentals of data science or working with typically large datasets and databases. Practitioners from industry, government, or research organisations with some basic training in quantitative analysis or computer programming are also welcome. Because this course surveys diverse techniques and methods, it makes an ideal foundation for more advanced or more specific training. Our applications are drawn from social, political, economic, legal, and business and marketing fields, rather than engineering or other sciences.

Objectives

This course aims to provide an introduction to the data science approach to the quantitative analysis of data using the methods of statistical learning, an approach blending classical statistical methods with recent advances in computational and machine learning. We will cover the main analytical methods from this field with hands-on applications using example datasets, so that students gain experience with and confidence in using the methods we cover. We also cover data preparation and processing, including working with structured databases, key-value formatted data (JSON), and unstructured textual data. At the end of this course students will have a sound understanding of the field of data science, the ability to analyse data using some of its main methods, and a solid foundation for more advanced or more specialised study.

The course will be delivered as a series of morning lectures, followed by lab sessions in the afternoon where students will apply the lessons in a series of instructor-guided exercises using data provided as part of the exercises. The course will cover the following topics:

- an overview of data science and the challenge of working with big data using statistical methods
- how to integrate the insights from data analytics into knowledge generation and decision-making
- how to acquire data, both structured and unstructured, and to process it, store it, and convert it into a format suitable for analysis
- the basics of statistical inference including probability and probability distributions, modelling, experimental design
- an overview of classification methods and related methods for assessing model fit and cross-validating predictive models
- supervised learning approaches, including linear and logistic regression, decision trees, and naive Bayes
- unsupervised learning approaches, including clustering, association rules, and principal components analysis
- quantitative methods of text analysis, including mining social media and other online resources

Prerequisites

An introduction to quantitative methods at any level would serve as a very useful foundation for this course, although no formal prerequisites are required. Familiarity with computer programming or database structures is a benefit, but not formally required.

Preparing before the course

We strongly recommend you spend some of July and August before the course reading some of the following materials:

- James et al (2013), Chapters 1–2
- *An Introduction to R*, available from https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Downloading and installing RStudio, available from http://www.rstudio.com.
- A brief on-line introduction to RMarkdown, which we will use for completing the exercises for the course, see https://goo.gl/ZqOwUe

Important Specifics

Computer Software

Computer-based exercises will feature prominently in the course, especially in the lab sessions. The use of all software tools will be explained in the sessions, including how to download and install them. This year we will be working primarily in R, using the `quanteda` package.

Main Texts

The primary texts are:

- James et al. (2013) An Introduction to Statistical Learning: With applications in R. Springer.
- Zumel, N. and Mount, J. (2014). Practical Data Science with R. Manning Publications.

The following are supplemental texts which you may also find useful:

- Lantz, B. (2013). Machine Learning with R. Packt Publishing.
- Conway, D. and White, J. (2012) Machine Learning for Hackers . O'Reilly Media.
- Leskovec, J., Rajaraman, A. and Ullman, J. (2011). Mining of Massive Datasets . Cambridge University Press.
- Zafarani, R., Abbasi, M. A. and Liu, H. (2014) Social Media Mining: An introduction . Cambridge University Press.

Instructors

Kenneth Benoit is Professor of Quantitative Social Research Methods at the Department of Methodology, LSE. With a background in political science, his substantive work focuses on political party competition, political measurement issues, and electoral systems. His research and teaching is primarily in the field of social science statistical applications. His recent work concerns the quantitative analysis of text as data, for which he has developed a package for the R statistical software.

Dr. Slava Mikhaylov is a Senior Lecturer in Quantitative Methods at UCL and has been teaching quantitative methods at UCL Political Science department for the last five years. HeâĂŹs currently involved in an ESRC Big Data infrastructure investment initiative âĂŞ Consumer Data Research Centre at UCL. One of SlavaâĂŹs responsibilities in the Centre is development and provision of big data analytics training for academic and professional community (data users). In addition Slava Mikhaylov is deputy director of UCL Q-Step Centre, an ESRC-funded initiative to promote quantitative methods.

Dr. Paul Nulty will serve as the lab assistant for this course.

Short Course Schedule

Day	Date	Topic(s)	Details
Mon	17 Aug	Course overview and introduction to data science	We will use this session to get to know the range of interests and experience students bring to the class, as well as to survey the approaches to be covered. We will also discuss and demonstrate the R software.
Tue	18 Aug	Research design issues in data science	Sampling, causal inference from observational data, differences with experimental settings, features. variables. Basic probability and statistics, binomial and Normal distributions. Cross-validation, predictive accuracy versus marginal effects. Model selection.
Wed	19 Aug	Linear Regression	The basic linear regression model, with a focus on prediction.
Thu	20 Aug	Generalized linear regression	Logistic regression, GAMs.
Fri	21 Aug	Resampling methods	Cross-validation, bootstrap.
Mon	24 Aug	Association rules and clustering	Cluster analysis, k-means clustering, and hierarchical clustering
Tue	25 Aug	Machine Learning	Decision trees, k-Nearest Neighbour, Naive Bayes. Evaluation metrics, precision and recall, cross-validation.
Wed	26 Aug	Unsupervised learning and dimensional reduction.	Principal components analysis, correspondence analysis.
Thu	27 Aug	Text analysis	Working with text in R, sentiment analysis, dictionary methods.
Fri	28 Aug	Mining the Social Web	Working with the Twitter API, Facebook API, JSON data, and examples.

Detailed Course Schedule

Monday, August 17: Overview and introduction to data science [KB, SM]

We will use this session to get to know the range of interests and experience students bring to the class, as well as to survey the approaches to be covered. We will also discuss and demonstrate the R software.

Required Reading:

- James et al (2013), Chapters 1–2.
- Zumel and Mount, Chapter 2.
- *An Introduction to R*, available from https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- Downloading and installing RStudio, available from http://www.rstudio.com.
- A brief on-line introduction to RMarkdown, which we will use for completing the exercises for the course, see https://goo.gl/ZqOwUe.

Recommended Reading:

- Patrick Burns, 2011. *The R Inferno*. Available from http://www.burns-stat.com/pages/Tutor/R_inferno.pdf.
- Lantz, Ch. 2.

Exercise 1: Getting started with R and RMarkdown.

This and subsequent exercises can be found from the course GitHub page.

Tuesday, August 18: Research design issues in data science [SM, KB]

This session will cover the topics of sampling, causal inference from observational data, differences with experimental settings, features, variables. We will discuss fundamental concepts of probability and statistics. This session will also introduce the topics of cross-validation, predictive accuracy versus marginal effects, and model selection.

Required Reading:

- James et al. Chapter 1-2.
- Zumel and Mount, Chapters 2-3.

Exercise 2: TBA

Wednesday, August 19: Linear Regression [SM]

This session will cover the basic linear regression model, with a focus on prediction.

Required Reading:

- James et al. Chapter 3.
- Zumel and Mount, Chapter 7.1.

Recommended Reading:

- Lantz, Chapter 6

Exercise 3: TBA

Thursday, August 20: Generalized Linear Regression [SM]

This session will cover logistic regression, generalized additive models (GAMs).

Required Reading:

- James et al. Chapters 4, 7.6–7.7.
- Zumel and Mount, Chapter 7.2.

Exercise 4: TBA

Friday, August 21: Resampling methods [SM]

This session will introduce resampling methods. We will cover cross-validation and bootstrapping.

Required Reading:

- James et al. Chapter 5.
- Zumel and Mount, Chapters 5-6.

Exercise 5: TBA

Monday, August 24: Association rules and clustering [KB]

This session will cover cluster analysis, k-means clustering, and hierarchical clustering.

Required Reading:

- James et al. Chapter 10.3
- Zumel and Mount, Chapter 8

Recommended Reading:

Lantz, Chapter 8

Exercise 6: TBA

Tuesday, August 25: Machine Learning [KB]

This session will cover decision trees, k-Nearest Neighbor, Naive Bayes. We will also explore evaluation metrics, precision and recall, and cross-validation.

Required Reading:

- James et al. Chapter 8
- Zumel and Mount, Chapter 5

Recommended Reading:

- Lantz, Chapter 10

Exercise 7: TBA

Wednesday, August 26: Unsupervised learning and dimensional reduction [KB]

This session will cover principal components analysis, correspondence analysis.

Required Reading:

- James et al. Chapter 10.1–10.2

Recommended Reading:

TBA

Exercise 8: TBA

Thursday, August 27: Text analysis [KB]

This session will cover working with text as data in R, sentiment analysis, dictionary methods.

Required Reading:

- Grimmer, J, and B M Stewart. 2013. "Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*
- Benoit, Kenneth and Alexander Herzog. In press. "Text Analysis: Estimating Policy Preferences From Written and Spoken Words." In Analytics, Policy and Governance, eds. Jennifer Bachner, Kathyrn Wagner Hill, and Benjamin Ginsberg. 21(3): 267âĂŞ97.

Recommended Reading:

TBA

Exercise 9: Analysing text using the `quanteda` package

Friday, August 28: Mining the Social Web [KB, SM]

This session will cover working with the Twitter API, Facebook API, JSON data.

Required Reading:

TBA

Recommended Reading:

TBA

Exercise 10: Working with Twitter data