CSCI E-63C: Elements of Data Science and Statistical Learning with R
Spring 2018

General Course Information

Instructor(s): Dr. Andrey Sivachenko, Dr. Victor Farutin
Office:
Office Hours:
Phone: 801-450-5934
E-mail: asivachenko@fas.harvard.edu, vif541@g.harvard.edu
Web site: https://canvas.harvard.edu/courses/34998
Classroom: Harvard Hall 104
Class Times: Thursday 7:40pm - 9:40pm

Prerequisites: (1) Familiarity with computer programming (CSCI-10a or equivalent); particular programming language is not important, being able to write a simple program and understanding program's flow control (conditionals, loops), variables and functions is. (2) Familiarity with basic probability and statistics STAT E-100 or equivalent): the concepts of marginal and conditional probability, random variable, probability density distribution, statistical testing, regression.

Note on prerequisites and pretest:

This course covers substantial amount of material in a short period of time - to do well students have to be comfortable with programming and the fundamentals of statistics (at least semester of instruction in each). We have prepared a short pretest to help you understand what level of familiarity with these two domains is anticipated for successful participation in the class. Please set aside some time to take the test. We won't use your score to keep you out of the class, but consider the results as you decide if this is the right course for you. This pretest is first and foremost for your own benefit to help decide whether you have sufficient background to get the most of what this course offers or whether the amount of material covered might present too steep of a learning curve.

The primary goal of the pretest is to help you understand whether your command of coding and stats positions you for a success in this class. Each of the ten questions in the test can be easily answered after a few minutes on google, but this is NOT the point: the working knowledge that would allow you to do well in this course is approximately equivalent to finding answers to these questions rather obvious upon careful reading and consideration of them on your own without external input. The workload for this course is designed under the assumption that this is the case for the vast majority of them. The pretest and guidelines for interpretation of the results can be found at: **https://harvard.az1.qualtrics.com/jfe/form/SV_eamCFWzGlH1olSd**

Textbook: *An Introduction to Statistical Learning with Applications in R*
[this book is also widely known and referred to simply as *ISL*]
Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Publisher: Springer (1st Ed 2013, corr 5th printing 2015)
ISBN-13: 978-1461471370
PDF version of the book can be downloaded from the author's homepage

Course Objectives

One of the broad goals of data science is examining raw data with the purpose of identifying its structure and trends, and of deriving conclusions and hypotheses from it. In the modern world awash with data, data analytics is more important than ever to fields ranging from biomedical research, space and weather science, finance, business operations and production, to marketing and social media applications.

This course introduces various statistical learning methods and their applications. The R programming language, a very popular and powerful platform for scientific and statistical analysis and visualization, is also introduced and used throughout the course. The fundamentals of statistical testing and learning are discussed, and the topics covered include linear and non-linear regression, clustering and classification, neural networks, support vector machines, and decision trees.

Specific topic coverage includes:

- Elementary probability and statistics (brief overview)
- Definition, principles and different types of statistical learning, model quality, bias-variance tradeoff
- Simple and multiple linear regression
- Assessing model prediction quality, cross-validation, bootstrap
- Model selection and regularization: dimensionality reduction, ridge and lasso
- Unsupervised learning: clustering approaches (K-means, hierarchical clustering)
- Supervised learning: classification problem
- Classification using logistic regression, naïve Bayes
- Classification with Support Vector Machines
- Neural Networks

Course information is available at **https://canvas.harvard.edu/courses/34998** This Web site contains class announcements and notes, test dates, lecture slides and videos, the course syllabus, and additional information.

If you have any questions about the course or need assistance, please contact course instructors in person or by telephone or by e-mail at any time, or you can also email the TAs assigned to the class.

All class assignments should be submitted via Canvas website on or before the due date. The solutions to the practical data analysis exercises *must be submitted in the form of both R markdown source file and HTML file generated from it.* The first week materials will include brief introduction to R markdown and compilation of analysis results in HTML format using Rstudio.

Grading and Evaluation Criteria

50% of the grade is based on the midterm and the final examination (25% each).

50% of the grade is based on homeworks (i.e. simple average of homework grades times 0.5). Homework consists of a quiz (40 points) and coding data analysis problems (60 points). Solutions for homework problems that ask for writing a program must include full working code in order to get full credit.

Late submission policy

This class involves substantial amount of homework and keeping up with the weekly assignments is essential for succeeding in this course. If you find out that you need an extension, please coordinate with instructors ahead of time that week, so that they are aware of it, otherwise submissions past deadline will result in losing substantial amount of points. Specifically, solutions for the assignments submitted later than 1, 2, 3 and 4 days after due date will be penalized respectively by 10%, 20%, 50% and 100% of the total points available for that assignment.

Course Outline

Week	Topics	Chapter Readings
1 Jan 25	Course introduction and overview of basic probability and statistics: statistical description of phenomena, statistical samples, sampling error; hypothesis testing and different statistical tests (parametric, non-parametric). R language: data types, vectors and matrices, indexing operators, functions, simple graphs; Rstudio, Rmarkdown	Chapter 1
2 Feb 1	Overview of basic probability and statistics – continued; Definition and overview of “statistical learning”, different types of tasks: supervised vs unsupervised learning, regression, classification; statistical models and model assessment: quality of fit, prediction accuracy. R language: lists, data frames (“tables”), loading data, installing packages	Chapter 2
3 Feb 8	Regression: K-Nearest Neighbors (KNN), simple linear regression, quality of fit, diagnostic plots; training and test error rates, model validation with introduction to cross-validation; R language: library functions for performing linear regression and model assessment in R	Chapters 3, 5
4 Feb 15	Regression continued: multiple linear regression, interaction terms, assessing model significance (anova, nested models, information criteria); resampling (cross-validation and bootstrap); R language: practicing regression, evaluating significance of the terms, cross-validation; visualizations of model fits and diagnostic plots.	Chapters 3, 5
5 Feb 22	Model selection and regularization: ridge regression, lasso, dimensionality reduction (PCA), stepwise/forward selection. Developing and examining examples of R code for model selection.	Chapter 6

6 **Mar 1**	Unsupervised learning: PCA, introduction to K-means clustering and hierarchical clustering	Chapter 10
7 **Mar 8**	**Midterm Exam. No lecture.**	
Mar 15	Spring break, NO LECTURE	
8 **Mar 22**	Unsupervised learning, continued: case studies, comparing different methods and approaches, distance metrics, assessing quality and robustness of clusters; developing and studying R code for performing clustering and generating visualizations	
9 **Mar 29**	Classification problem; naïve Bayes classifier, classification with KNN, logistic regression; R language practice: using R for classification tasks	Chapter 4
10 **Apr 5**	Classification, continued: tree-based methods (“decision trees”); model ensembles, boosting, bagging and random forests; using R for decision-tree based classification.	Chapter 8
11 **Apr 12**	Support vector machines (SVM): maximal margin, support vector classifiers, relationship to logistic regression; R language: exploring R libraries for SVM-based classification	Chapter 9
12 **Apr 19**	Classification, continued: neural networks (NN), assessing and comparing performance of different classification models and algorithms, ROC curves	
13 **Apr 26**	Case study: end to end analysis of a multi-dimensional real life dataset	
14 **May 3**	Review session	
15 **May 10**	**Final Exam. No lecture.**	

Accessibility

The Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit www.extension.harvard.edu/resources-policies/resources/disability-services-accessibility for more information.

Cheating and plagiarism

You are responsible for understanding Harvard Extension School policies on academic integrity (www.extension.harvard.edu/resources-policies/student-conduct/academic-integrity) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity. To support your learning about academic citation rules, please visit the Harvard Extension School Tips to Avoid Plagiarism (www.extension.harvard.edu/resources-policies/resources/tips-avoid-plagiarism), where you'll find links to the Harvard Guide to Using Sources and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.