GRS MA 751 -- Statistical Machine Learning

Boston University, Spring 2022

Instructor: Professor Mark Kon
Office: MCS 260
Phone: 617-353-9549
Email: mkon@bu.edu
Course Website: http://math.bu.edu/people/mkon/ (navigate to courses)
Office Hours: Tu 3:30-4:45 pm; Th 6:30-7:45 pm
Class Meetings: TR 5:00 pm-6:15 pm.

Course Description: This course studies fundamental principles of machine learning from a statistical standpoint. The goal is to look 'under the hood' at the variety of methods used to learn from data. An important underlying principle is the large number of universals in this area: many seemingly different algorithms are in fact manifestations of identical or similar principles. It is important to understand that understanding just the procedure or pseudocode of an algorithm is in most cases merely scratching its surface, and does not constitute genuine understanding.

We will cover a selection of topics in statistical learning, such as regularized basis methods, kernel methods, boosting, neural networks, support vector machines, and graphical models.

It is also possible to study machine learning from a computer science viewpoint (more algorithmically); you can see such material in the Stanford course of Andrew Ng (https://see.stanford.edu/Course/CS229) as well as resources there and elsewhere

There are also fundamental approaches using more mathematical standpoints, which emphasize methods largely in functional analysis. Mathematical aspects of kernel methods, reproducing kernel Hilbert spaces and tools such as VC dimension can be studied in more detail in such courses; see the class of Tomaso Poggio (Statistical Learning Theory and Applications) at MIT (https://cbmm.mit.edu/9-520); this course will be offered in the Fall of 2022.

Prerequisites: CAS MA 581 (Probability) or equivalent, MA 575 (Linear Models) or equivalent course on regression methods; or consent of instructor.

Text: T. Hastie, R. Tibshirani, and J. Friedman. Elements of Statistical Learning: Data Mining, Inference, and Prediction (Second Edition). Springer; New York. 2009.

Software: Instruction will be done using the statistical software R for computing in this course. The software is free and compatible with Windows, Mac, and Linux/Unix. It may be downloaded from cran.us.r-project.org. The main website for the R project is www.r-project.org. For those who do not wish to install R on their own machines, it is possible to access R (as well as Matlab and other software) on the BU Linux Virtual Lab system. You can use R and Matlab on the campus Linux Virtual Lab system under Linux

using XWindows to access the Linux Virtual Lab. This is a part of the SCC (Shared Computing Cluster) at BU. For information on Linux Virtual Lab see http://www.bu.edu/tech/services/support/desktop/computer-labs/unix/ , and specific information on running Matlab is at http://www.bu.edu/tech/support/research/software-and-programming/common-languages/matlab/ . From off campus you will need a VPN connection to do this (see http://www.bu.edu/tech/services/support/remote/vpn/). You can also purchase the student edition of Matlab for use on your own Mac or PC at the B.U. Barnes and Noble bookstore (the cost, however, is approximately $50, and more for the student version on Amazon). This will give you a code for downloading the software to your computer. You can also use the BU PC Lab at the BU Common at Mugar (see http://www.bu.edu/common/). If you are using Matlab on the BU SCC, make sure you are in X-Windows mode, and not just terminal mode; you can also access Xwindows from a PC or Mac as mentioned above. For those students in the department, R is also available on the machines in the Department of Mathematics and Statistics.

Use of R is not required for the course. If you prefer to use a different package or programming language/environment, you are free to do so. But you will be responsible for seeing that you have sufficient access to software tools for the various topics covered in the course.

Course Format: Much of the material in this course involves developments in the field of statistical machine learning from just the past 15 years or so. And many topics are from areas still under active development. The goal of the course is to both serve as an introduction to the underlying problems and principles and key methodologies in this field and to develop an improved facility for quickly and efficiently navigating the mix of analytical and computational aspects necessary for ?getting up to speed? on approaches in this area.

We will aim to cover roughly a chapter or two per week. Corresponding to each chapter will be (i) reading from the textbook, (ii) analytical completion problems, and (iii) data analyses. In addition, there will be a final course project, and a final exam.

1. Completion Problems: Problems from the end of each chapter will be assigned as problem sets, which will be due on Thursdays. Students are encouraged to work together on these problems, though what is handed in must be done individually.

2. Data Analyses: At a spacing of roughly every few weeks (depending on the particular chapters), a data analysis will be assigned, asking students to implement and/or explore the tools covered during a certain portion of the course. These assignments will involve a somewhat more substantial amount of work than the completion problems, and will be handed in separately and graded.

3. Course Project: The course project will ask you to choose a recently introduced methodology from the literature, one not covered during the course itself, and conduct a thorough investigation of that methodology. Project proposals will need to be approved by the course instructor. Project results will be turned in in the form of a written report,

including a summary of background and the proposed methodology, implementation details, and description of all simulations, analyses, etc. Further details will be made available later in the semester.

4. Exams:
(a) There will be an 'in-class' midterm exam for the course. It will be given during a class hour.
(b) There will be an 'in-class' final exam based on course-related material, which will be given at the end of the semester.

5. Grading: The final grade for the course will be determined according to the following formula. 25% of the grade will be determined by analytical problems, 15% by data analyses, 25% by the final project, 15% by the midterm, 20% by the final exam.

6. Collaborative work: Learning is better done with others, and collaboration and discussion of problem sets in this course is encouraged. The purpose of this is to exchange ideas and methods conceptually, but not to 'copy solutions' from others. The final copies of problem solutions must be your own; direct copying/transcription of others' work is not permitted and is considered to be plagiarism by this and all other University codes of ethics (see below).

Please Note: Students are responsible for knowing, and abiding by, the provisions of the University Academic Conduct Code, which is posted at
https://www.bu.edu/academics/policies/academic-conduct-code/
Violations of the code are punishable by sanctions including expulsion from the University.

Course Syllabus (approximate):

1. **Week 1:** Chapter 2 / Chapter 3
2. **Week 2:** Chapter 3 continued
3. **Week 3:** Chapter 4
4. **Week 4:** Chapter 5
5. **Week 5:** Chapter 7
6. **Week 6:** Chapter 8
7. **Week 7:** Chapter 9
8. **Week 8:** Chapter 10
9. **Week 9:** Chapter 15 / Chapter 16
10. **Week 10:** Chapter 11
11. **Week 11:** Chapter 12
12. **Week 12:** Chapter 13
13. **Week 13:** Chapter 14
14. **Week 14:** Chapter 17
15. **Week 15:** Wrap-up of course