Syllabus for BST 263 Applied Machine Learning
Harvard T.H. Chan School of Public Health
Spring 2018

Course website:
https://canvas.harvard.edu/courses/34483

Instructor:
Jeffrey W. Miller
Assistant Professor of Biostatistics
Building 1 Room 419, 655 Huntington Ave, Boston MA 02115
jwmiller@hsph.harvard.edu

Teaching assistants:
Lin Liu
PhD Student, Biostatistics
linliu@fas.harvard.edu

Shayna Stein
PhD Student, Biostatistics
sstein@g.harvard.edu

Class time and location:
Tuesdays and Thursdays, 3:45-5:15 p.m., FXB G12

Office hours:
Dr. Miller: Tuesdays, 5:30-6:30 p.m., Building 2, room 426
Lin Liu: Thursdays, 5:30-6:30 p.m., Building 2, room 428
Shayna Stein: Mondays, 5:30-6:30 p.m., Building 2, room 434

Prerequisites:
This course is intended for students in all departments of the School. Students are required to have taken one of the following courses: BST 260 or BST 210 or BST 232, or an equivalent course.

Purpose of the course
Statistical machine learning is a collection of flexible tools and techniques for using data to construct algorithms for prediction and exploratory analysis. This 5-credit course will take place in the second semester of the Health Data Science Master's program, and will introduce students to the most essential elements of machine learning.

The central theme of the course will be to ground the material in practical real-world data examples, in order to motivate the concepts and illustrate why and how the methods work. Some mathematical foundations will be covered, but the primary emphasis of the course will be on learning how to implement and use the methods, while gaining an intuitive understanding of them. Programming (in R) will be used throughout the course to provide hands-on training.

Course structure

The course will consist of a series of modules, building up from foundations, through linear, non-linear, and nonparametric supervised methods, as well as covering unsupervised learning and Bayesian methods.

Along the way, the students will learn how to evaluate model performance/accuracy, quantify uncertainty, and combine methods via ensembles. The students will gain hands-on experience implementing and applying the methods in lab exercises and homework programming assignments, while learning the conceptual foundations in homework problem sets. There will be in-class labs in addition to lectures. Student performance will be assessed via homework assignments, a midterm exam, and a final exam.

In addition to homework assignments and exams, class attendance and thoughtful participation are important and will be reflected in part in the final grade. Informed student participation in classroom discussions is required of all students. Students are expected to behave professionally at all times, with courtesy towards other students, the TAs, and the instructor.

Course materials

Electronic copies of course notes/slides, homework assignments, and data sets will be posted on the course website. The textbook for the course is:

> (**ISL**) Gareth James, Daniella Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning", Springer Texts in Statistics. Electronic copy available for free at: http://www-bcf.usc.edu/~gareth/ISL/

For supplementary reading, a more in-depth treatment of similar material is provided in:

> Trevor Hastie, Robert Tibshirani, Jerome Friedman. "Elements of Statistical Learning", Springer Texts in Statistics. Electronic copy available for free at: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Grades and performance evaluation

Overall grades will be based on the following:

50% Homework

Homework assignments will consist of problem sets and computer programming assignments. The purpose of the homework assignments is to enable the students to more fully and deeply understand the concepts of the course, to gain practical experience implementing and using the methods introduced in the course, and to receive feedback on their performance and their understanding of the material.

20% Midterm exam

25% Final exam

There will be one midterm exam and one final exam. The exams will consist of problems similar to those encountered by students in the homework assignments and the in-class exercises. The purpose of the exams is to evaluate the students' understanding of the material, and provide feedback on their performance.

Exams will be closed book, closed notes. Exam dates will be announced far in advance, and there will be no make-up exams. Exams will be graded and returned within one week.

5% Class participation

Class attendance and thoughtful participation are important and will be reflected in part in the final grade. Informed student participation in classroom discussions is required of all students.

Homework policies

Due dates will be posted and homework submission is mandatory. Graded homework will be returned one week after submission.

Homework must be submitted online on the course website. Any handwritten material must be legible, otherwise no credit will be given. Kresge LL-19 and Countway Library have scanners that can be used at no cost. For programming exercises, include (a) plots and numerical results when appropriate, (b) discussion of the results when appropriate, (c) any supporting derivations, written out separately from the code, and (d) your source code (typed). The TAs will not run your code (e.g., to generate plots, etc.), so anything you want them to see must be included.

Policy on homework collaboration:

- Each student is required to come up with their own solutions for the homework.
- Students are allowed to discuss the problems in general terms (without sharing complete solutions) among themselves, or with the TAs or instructor. HOWEVER, when writing up their solutions, students are required to do this on their own, without copying from another source.
- Students are forbidden from using solutions from any other source (such as solutions found online).
- Violation of this policy will result in a score of zero for that assignment.

Late submission policy:

- Homework submissions will be timestamped, and late submissions will be penalized as follows: starting from the due time until 24 hours after the due time, a multiplicative penalty starting at 1.0 and decreasing linearly to 0.0 will be applied. So, for example, an assignment submitted 6 hours late will incur a penalty of 0.75 (75% credit), an assignment submitted 12 hours late will incur a penalty of 0.50 (50% credit), and an assignment submitted 24 hours late or later will incur a penalty of 0.0 (no credit).
- There will be no make-ups or extensions.

Course outline (tentative)

1. Introduction

Lecture 1: Introduction, Notation and probability basics

Reading: ISL 1

2. Statistical Learning (Overview)

Lecture 2: What is Statistical Learning?

Reading: ISL 2.1

Lecture 3: Assessing model accuracy, R programming basics

Reading: ISL 2.2-2.3

3. Linear regression

Lecture 4: Simple linear regression

Reading: ISL 3.1

Lecture 5: Multiple linear regression

Reading: ISL 3.2-3.3

Lecture 6: Illustrations, Lab on linear regression

Reading: ISL 3.4-3.6

4. Classification

Lecture 7: Classification overview, Logistic regression

Reading: ISL 4.1-4.3

Lecture 8: Lab on classification

Reading: ISL 4.6

5. Resampling methods

Lecture 9: Cross-validation and bootstrap

Reading: ISL 5.1-5.2

Lecture 10: Lab on cross-validation and bootstrap

Reading: ISL 5.3

6. Linear model selection and regularization

Lecture 11: Subset selection, Penalty-based methods

Reading: ISL 6.1-6.2

Lecture 12: Lab on subset selection, ridge regression, and Lasso

Reading: ISL 6.5-6.6

Lecture 13: Dimensional reduction, High-dimensional issues

Reading: ISL 6.3-6.4

Lecture 14: Lab on dimensional reduction

Reading: ISL 6.7

Midterm exam – Tentative date: March 22, 2018, at the usual class time and location.

If you have a conflict with this time/date, let me know by January 29. After January 29, the date will be set and there will be no makeup exams. If you absolutely must miss the midterm due to extraordinary circumstances, the weight given to your final exam will increase accordingly, so that you have the opportunity to make up the points.

7. Non-linear regression

Lecture 15: Polynomial regression, step functions, and basis functions

Reading: ISL 7.1-7.3

Lecture 16: Local regression, Generalized additive models (GAMs)

Reading: ISL 7.6-7.7

Lecture 17: Lab on non-linear modeling

Reading: ISL 7.8

8. Tree-based methods and ensembles

Lecture 18: Classification and regression trees (CART), bagging, and random forests

Reading: ISL 8.1-8.2

Lecture 19: Lab on CART and random forests

Reading: ISL 8.3

Lecture 20: Boosting, gradient boosting, and other ensemble methods

Reading: ISL 8.2, course notes

Lecture 21: Lab on ensemble methods

Reading: Course notes

9. Support vector machines

Lecture 22: Maximum margin classifier, Support vector classifier

Reading: ISL 9.1-9.2

Lecture 23: Kernel trick, Support vector machine, Multi-class problems

Reading: ISL 9.3-9.4

Lecture 24: Lab on support vector machines

Reading: ISL 9.6

10. Unsupervised learning

Lecture 25: PCA review, K-means and hierarchical clustering

Reading: ISL 10.1-10.3

Lecture 26: Lab on clustering

Reading: ISL 10.5-10.6

11. Bayesian methods

Lecture 27: Fundamentals of Bayesian statistics

Reading: Course notes

Lecture 28: Gibbs sampling

Reading: Course notes

Lecture 29: Lab on Gibbs sampling

Reading: Course notes

Final exam – Time/place to be determined

Harvard Chan Policies and Expectations

Inclusivity Statement
Diversity and inclusiveness are fundamental to public health education and practice. It is a requirement that you have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Academic Integrity
Harvard University provides students with clear guidelines regarding academic standards and behavior. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great

care to distinguish their own ideas and knowledge from information derived from sources. Please refer to policy in the student handbook for details on attributing credit and for doing independent work when required by the instructor.

Accommodations for Students with Disabilities
Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact the OSA studentaffairs@hsph.harvard.edu in all cases, including temporary disabilities.

Course Evaluations
Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement. Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.