Syllabus for BST 263 Statistical Learning

Harvard T.H. Chan School of Public Health
Spring 2018

Course website:
https://canvas.harvard.edu/courses/55674

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

Teaching assistants:
Yuri Ahuja, PhD Student, Biostatistics
yuri_ahuja@hms.harvard.edu

Kareem Carr, PhD Student, Biostatistics
kareemcarr@g.harvard.edu

Greyson Liu, PhD Student, Biostatistics
gang_liu@g.harvard.edu

Class time and location:
Tuesdays and Thursdays, 3:45-5:15 p.m., Kresge G2

Office hours:
Greyson – Tuesdays, 1:00-2:00 p.m., Kresge LL6
Yuri – Wednesdays, 5:30-6:30 p.m., Building 2, room 434
Dr. Miller – Thursdays, 2:00-3:00 p.m., Building 2, room 434
Kareem – Fridays, 1:30-2:30 p.m., Building 2, room 428

Prerequisites:
Students are required to have taken one of the following: BST 260 or BST 210 or BST 232.

Purpose of the course

Statistical learning is a collection of flexible tools and techniques for using data to construct prediction algorithms and perform exploratory analysis. This course will introduce students to the theory and application of methods for supervised learning (classification and regression) and unsupervised learning (dimension reduction and clustering). Students will learn the mathematical foundations underlying the methods, as well as how and when to apply different methods. Topics will include the bias-variance tradeoff, cross-validation, linear regression, logistic regression, KNN, LDA/QDA, variable selection, penalized regression, generalized additive models, CART, random forests, gradient boosting, kernels, SVMs, PCA, and K-means. Homework will involve mathematical and programming exercises, and exams will contain conceptual and mathematical problems. Programming in R will be used throughout the course to provide hands-on training and practical examples.

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 unsupervised learning methods.

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. Labs count toward class participation and bonuses will be awarded based on performance in lab competitions. Grades will primarily be determined by homework assignments, the midterm exam, and the final exam.

Course materials

Electronic copies of lecture slides, videos of the lectures, homework assignments, labs, 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:

> (**ESL**) 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 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 finalized 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 labs and 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.

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. However, do NOT submit excessively long homework submissions with pages and pages of detailed computer output. Points will be taken off for excessively long submissions that unnecessarily burden the TAs.

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, and possible disciplinary action.

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).
- Late lab submissions receive zero credit and are ineligible for bonuses.
- There will be no make-ups or extensions.

Course outline / schedule of topics by class period

Introduction and background

1. (Jan 29) Course overview, Choosing among methods

   Reading: Slides, ISL 1

2. (Jan 31) Probability and linear algebra basics

   Reading: Slides

Performance tradeoffs

3. (Feb 5) Measuring performance, Bias-variance tradeoff, KNN classifier

   Reading: Slides, ISL 2.1-2.3

4. (Feb 7) Lab on KNN

Linear regression

5. (Feb 12) Linear regression

   Reading: Slides, ISL 3.1-3.5

6. (Feb 14) Lab on linear regression and bias-variance tradeoff

Classification

7. (Feb 19) Classification, Logistic regression, LDA/QDA

   Reading: Slides, ISL 4.1-4.5

8. (Feb 21) Lab on classification

Cross-validation

9. (Feb 26) Cross-validation

   Reading: Slides, ISL 5.1, 5.3

10. (Feb 28) Lab on cross-validation

Penalized regularization

11. (Mar 5) Ridge regression, Lasso, Subset selection, Model selection

    Reading: Slides, ISL 6.1-6.2

12. (Mar 7) Lab on penalized regression

Principal components analysis

13. (Mar 12) Principal components analysis (PCA)

Reading: Slides, ISL 6.3, ISL 10.1-10.2

14. (Mar 14) Lab on PCA

Midterm

15. (Mar 26) Lab on the Caret package

16. (Mar 28) Midterm exam

    Usual class time and location. Contact me by Feb 5 if you have a religious holiday that conflicts with this date. Otherwise, if you miss the midterm exam, the weight given to your final exam will increase to 45% instead of 25%.

Nonlinear methods

17. (Apr 2) Splines, smoothing splines

    Reading: Slides, ISL 7.1-7.5

18. (Apr 4) Local regression, Generalized additive models (GAMs)

    Reading: Slides, ISL 7.6-7.7

19. (Apr 9) Lab on nonlinear methods

Trees

20. (Apr 11) Classification and regression trees (CART)

    Reading: Slides, ISL 8.1

21. (Apr 16) Bagging and Random forests

    Reading: Slides, ISL 8.2

22. (Apr 18) Lab on trees

Boosting

23. (Apr 23) Boosting and Gradient boosting

    Reading: Slides, ISL 8.2

24. (Apr 25) Lab on boosting

Support vector machines

25. (Apr 30) Maximum margin classifier, Support vector classifier

    Reading: Slides, ISL 9.1-9.2

26. (May 2) Kernel trick, Support vector machine, Multi-class problems

    Reading: Slides, ISL 9.3-9.4

Clustering

27. (May 7) K-means and Hierarchical clustering

    Reading: Slides, ISL 10.3

28. (May 9) Lab on clustering

Final

29. (May 14) Review session

30. (May 16) Final exam

    Usual class time and location. Contact me by Feb 5 if you have a religious holiday that conflicts with this date. Otherwise, there will be no makeup exam, so make sure you are free on this date.

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.