

# 36-708 Statistical Methods in Machine Learning

Syllabus, Spring 2019

<http://www.stat.cmu.edu/~larry/=sml>

**Lectures:** Tuesday and Thursday 1:30 - 2:50 pm (POS 152)

This course is an introduction to Statistical Machine Learning. The goal is to study modern methods and the underlying theory for those methods. There are two pre-requisites for this course:

1. 36-705 (Intermediate Statistical Theory)
2. 10-707 (Regression)

## Contact Information

### Instructor:

Larry Wasserman BH 132G 412-268-8727 [larry@cmu.edu](mailto:larry@cmu.edu)

### Teaching Assistants:

The names and office hours for the TA's will be on the course website.

### Office Hours

Larry Wasserman Tuesdays 12:00-1:00 pm Baker Hall 132G

## Text

There is no text but course notes will be posted. Useful reference are:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman (2001). *The Elements of Statistical Learning*, Available at <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>.
2. Chris Bishop (2006). *Pattern Recognition and Machine Learning*.
3. Luc Devroye, László Györfi, Gábor Lugosi. (1996). *A probabilistic theory of pattern recognition*.
4. Györfi, Kohler, Krzyzak and Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*.
5. Larry Wasserman (2004). *All of Statistics: A Concise Course in Statistical Inference*.
6. Larry Wasserman (2005). *All of Nonparametric Statistics*.

## Grading

1. There will be four assignments. They are due **Fridays at 3:00 p.m.** Hand them by uploading a pdf file to Canvas.
2. **Midterm Exam.** The date is **Thursday MARCH 7**.
3. **Project.** There will be a final project, described later in the syllabus.

Grading will be as follows:

**50% Assignments**  
**25% Midterm**  
**25% Project**

## Policy on Collaboration

Collaboration on homework assignments with fellow students is encouraged. However, such collaboration should be clearly acknowledged, by listing the names of the students with whom you have had discussions concerning your solution. You may not, however, share written work or code after discussing a problem with others. The solutions should be written by you.

## Topics

1. Introduction and Review
  - (a) Statistics versus ML
  - (b) concentration
  - (c) bias and variance
  - (d) minimax
  - (e) linear regression
  - (f) linear classification
  - (g) logistic regression
2. Nonparametric Inference
  - (a) Density Estimation
  - (b) Nonparametric Regression Regression
    - i. kernels
    - ii. local polynomial
    - iii. NN
    - iv. RKHS
  - (c) Nonparametric Classification
    - i. plug-in
    - ii. nn
    - iii. density-based
    - iv. kernelized SVM
    - v. trees
    - vi. random forests
3. High Dimensional Methods
  - (a) Forward stepwise regression
  - (b) Lasso
  - (c) Ridge Regression
  - (d) High dimensional classification
4. Clustering
5. Graphical models
6. Minimax theory
7. Causality
8. Dimension reduction: PCA, nonlinear
9. Other Possible Topics
  - (a) Mixture models
  - (b) Wasserstein distance and optimal transport
  - (c) boosting
  - (d) active learning

- (e) nonparametric bayes
- (f) deep learning
- (g) differential privacy
- (h) interactive data analysis
- (i) multinomials
- (j) statistics computational tradeoff
- (k) random matrices
- (l) conformal methods
- (m) interpolation

## Course Calendar

The course calendar is posted on the course website and will be updated throughout the semester.

## Project

The project involves picking a topic of interest, reading the relevant results in the area and then writing a short paper (8 pages) summarizing the key ideas in the area. You may focus on a single paper if you prefer. You are NOT required to do new research.

The paper should include background, statement of important results, and brief proof outlines for the results. If appropriate, you should also include numerical experiments or an application with real data.

1. You may work by yourself or in teams of two.
2. The goals are (i) to summarize key results in literature on a particular topic **and** (ii) present a summary of the theoretical analysis (results and proof sketch) of the methods (iii) implement some of the main methods. You may develop new theory if you like but it is not required.
3. You will provide: (i) a proposal, (ii) a progress report and (iii) a final report.
4. The reports should be well-written.

**Proposal.** A one page proposal is due **February 8**. It should contain the following information: (1) project title, (2) team members, (3) precise description of the problem you are studying, (4) anticipated scope of the project, and (5) reading list. (Papers you will need to read).

**Progress Report.** Due **April 5**. Three pages. Include: (i) a high quality introduction, (ii) what have you done so far, (iii) what remains to be done and (iv) a clear description of the division of work among teammates, if applicable.

**Final Report:** Due **May 3**. The paper should be in NIPS format. (pdf only). **Maximum 8 pages**. No appendix is allowed. You should submit a pdf file electronically. It should have the following format:

1. Introduction. Motivation and a quick summary of the area.
2. Notation and Assumptions.
3. Key Results.
4. Proof outlines for the results.
5. Implementation (simulations or real data example.)
6. Conclusion. This includes comments on the meaning of the results and open questions.