Statistics 154 Syllabus (14 weeks), Spring 2019

Modern Statistical Prediction and Machine Learning

Instructor: Professor Bin Yu (binyu@berkeley.edu);
office hours: Tu: 9:30-10:30; Wed: 9-10, in 409 Evans
GSIs: Yuansi Chen and Raaz Dwivedi (offce hours to be announced)
yuansi.chen@berkeley.edu; raaz.rsk@berkeley.edu

Logistics
Lecture: TU, TH 8:00 am - 9:29 am, VSL 2040
Section: Two hours / week
Homework: biweekly (including math and simulation parts)
Projects: 2 projects total, in groups
One midterm + One final

Grading
5% attendance + 20% final + 20% midterm + 20% project + 30% homework

Policy
No late homework or late project, no make-up midterm or make-up final

In-class midterm on March 21 (Thurs), Final exam on May 16, 7-10 pm

Code of conduct
No-copying homework or project; discussions are encouraged with required credits in submitted work to attribute to other's efforts
GSIs are trained to detect misconducts, esp. our GSIs having a proven record.

Course description

This course aims at training students to solve real world prediction problems. We achieve our goal through data experience and training of critical thinking, use of domain knowledge, data visualization, machine learning algorithms and mathematics. We take a holistic view of prediction in the data science life cycle that consists of problem formulation, data collection, exploratory data analysis (EDA), unsupervised learning, supervised learning, data results, validation, and conclusion.

We separate the real world from the world of mathematics and algorithm, and cover systematic methods for seeking evidence to connect the two worlds (sometimes well but often not). In particular, to help connect the two worlds, we emphasize the concepts PQR-S: P for population, Q for question, R for representativeness and S for scrutiny, and we follow the PCS framework: P for predictability, C for computability and S for Stability.

Attendance in class is necessary to do well in the class

Textbooks
An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics) [2017 Edition] (required)
Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Statistical Models: Theory and Practice [2009 Edition] (recommended)
Author: David Freedman

Tentative Lecture schedule

- Introduction and quick review of math concepts (matrix, gradient descent, probability theory)
  [1/2 week] (1/22)
- Problem formulation with a case study and data collection
  [1/2 week] (1/24)
- Introduction to PCS, internal validity, Cross-validation (CV), PQR-S, and external validity
  [1 week] (1/29, 1/31)
- Data preprocessing, exploratory data analysis and unsupervised learning (ggplot, superheat, SVD, PCA, hierarchical clustering, k-means, EM)
  [2 weeks] (2/5, 2/7, 2/12, 2/14) LAB 1 OUT
- Least squares (LS), linear regression and regularizations/model selection
  [3 weeks] (2/19, 2/21, 2/26, 2/28, 3/5, 3/7)
- Model assessment and bias-variance trade-off
  [1/2 week] (3/12)
- Classification, linear classification (logistic, SVM)
  [2 weeks] (3/14, 3/19)
- Midterm in class (3/21)
- **Spring break** (3/25, 3/27)
- Classification, linear classification (logistic, SVM)
  [2 weeks] (4/2, 4/4) LAB 2 OUT
- Kernel methods [1 week] (4/9, 4/11)
- Nearest Neighbor and Tree-based methods
  [1 week] (4/16, 4/18)
- Boosting algorithms
  [1/2 week] (4/23)
- Neural nets
  [1/2 week] (4/25)
- Interpretable machine learning
  [1/2 week] (4/30)
- Final review (5/2)
- Final exam location TBA

16	Thurs	5/16/19	7–10 pm