PB HLTH C240C/STAT C245C
Biostatistical Methods: Computational Statistics with Applications in Biology and Medicine I
Sandrine Dudoit
Fall 2018

Syllabus

PB HLTH C240C–D/STAT C245C–D, Biostatistical Methods: Computational Statistics with Applications in Biology and Medicine I and II, both concern statistical methods and software for addressing questions that arise in current biological and medical research.

Neither course is a prerequisite for the other.

A common thread among the statistical questions discussed in these courses is their high-dimensional and computer-intensive nature.

The courses also concern statistical computing and computationally reproducible research, with emphasis on the R language and environment [www.r-project.org](http://www.r-project.org).

The statistical methods and software are motivated by and illustrated on data generated by state-of-the-art biological assays.

Topics of interest, to be discussed in terms of both statistical methodology and software implementation, include the following.

- Numerical and graphical summaries of data.
- Dimensionality reduction: Principal component analysis (PCA), multidimensional scaling (MDS), biplots.
- Cluster analysis.
- Loss-based estimation: Parametric and non-parametric density estimation and regression (e.g., maximum likelihood estimation, linear regression, class prediction).
- Regression: Linear regression, generalized linear regression (GLM), non-linear regression, classification and regression trees (CART), nearest neighbor regression, linear and quadratic discriminant analysis, support vector machines (SVM).
- Smoothing: Robust local regression (lowess, loess), kernel density estimation, splines, multivariate adaptive regression splines (MARS).
- Regularization: Ridge regression, least absolute shrinkage and selection operator (LASSO), least angle regression (LARS), support vector machines (SVM).
- The expectation-maximization (EM) algorithm.
- Cross-validation.
• Loss-based estimation with cross-validation: Model/variable selection, performance/risk assessment.
• The bootstrap.
• Ensemble methods: Stacking, bagging, and boosting.
• Monte-Carlo methods: Markov chain Monte-Carlo (MCMC), importance sampling.
• Stochastic processes: Markov models, hidden Markov models (HMM).
• Multiple hypothesis testing.
• Dynamic programming.
• The design of in silico experiments.
• Computationally reproducible research.

Practical Matters

• Faculty instructor.
  Sandrine Dudoit
  Website: www.stat.berkeley.edu/~sandrine
  E-mail: sandrine@stat.berkeley.edu
  Office hours: Thursday, 14:00–15:00, 327 Evans Hall

• Graduate student instructor.
  Kelly Street
  Website: https://statistics.berkeley.edu/people/kelly-street
  E-mail: kstreet@berkeley.edu
  Office hours: TBD

• Time and location.
  Lecture: Tuesday and Thursday, 12:30–14:00, 330 Evans Hall
  Discussion: Wednesday, 12:00–14:00, 56 Barrows Hall

• Registration information.
  Public Health C240C, Class # 34277
  Statistics C245C, Class # 34324
  Units: 4

• Grading policy.
  50% assignments; 40% final project; 10% participation in lecture and discussion.
  Assignments involve both theory and data analysis using R and possibly other software.
  The final project consists of an abstract/proposal, written report, and poster or oral presentation on a
  topic that involves the application of statistical methods and software to address a particular biological
  or medical question.

  N.B. Attendance of the discussion is strongly encouraged, as 10% of the final grade is
  based on participation in both the lecture and discussion.

• References.
  There is no required textbook. Lecture notes and references will be provided on the class website.
• **Prerequisites.**

  **Statistics.** STAT 200A–B (may be taken concurrently) or STAT 201A–B (may be taken concurrently) or consent of instructor.

  **Computing.** Some familiarity with the R language. Tutorials are available on the R Project website ([www.r-project.org](http://www.r-project.org)) and on the UC Berkeley Statistical Computing Facility website ([http://statistics.berkeley.edu](http://statistics.berkeley.edu)); references are posted on the class website.

  **Biology.** No formal training in biology is required; basic notions will be presented in class and references will be provided for further learning.

  **N.B.** Please contact instructor if you do not satisfy the prerequisites. You are solely responsible for making up for any gaps in training.