PB HLTH C240D/STAT C245D
Biostatistical Methods: Computational Statistics with Applications in Biology and Medicine II

Sandrine Dudoit

Fall 2017

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 inference problems that arise in current biological and medical research.

Neither course is a prerequisite for the other.

A common thread among the statistical inference 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`).

The statistical methods and software are motivated by and illustrated on data structures that arise in current biological and medical research.

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), bi-plots.
- 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.
- Graph theory.
- 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: Tuesday, 14:00–15:00, 109 Haviland Hall
- *Graduate student instructor.*
  *** TBD
  Website: TBD
  E-mail: TBD
  Office hours: TBD
- *Time and location.*
  Lecture: Tuesday and Thursday, 12:30–14:00, 330 Evans Hall
  Discussion: Wednesday, 12:00–14:00, 342 Evans Hall
- *Registration information.*
  Public Health C240D, Class # 46094
  Statistics C245D, Class # 42738
  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 201A–B (may be taken concurrently) or old version STAT 200A–B or consent of instructor.

  Computing. Familiarity with the R language. Tutorials are available on the R Project website (`www.r-project.org`) and on the UC Berkeley Statistical Computing Facilty website (`http://statistics.berkeley.e` 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.