Big Data with Applications in Biostatistics

Instructor: Hongzhe Li
Fall 2020

This course covers topics from public heath and biomedical research where Big data are being collected and methods are being developed and applied, together with some core statistical methods in high dimensional data analysis.

Topic 0-: Introduction
(1) introduction to big data; (2) general statistical concepts and principles; (3) bias-variance trade in prediction.

Topic 1: Dimension reduction
(1) SVD and principal components analysis; (2) positive components analysis; (3) Applications in Genomics and integrative genomics. (4) Application to Netflix movie recommendation data. (5) Stochastic gradient descent

Topic 2: Unsupervised learning
Various clustering methods and GAP statistic, robust hierarchical clustering, variable selection for clustering analysis

Topic 3: Regularization and High Dimensional Regression Analysis
(1) general form of loss + regularization; (2) L2 regularization;
(3) L1 regularization and its variants (Lasso; adaptive Lasso;
elastic net; (4) Theory of Lasso; (5) Convex optimization, ADMM algorithm.
(6) Application to Google flu track

Topic 4: Ensemble Learning and Prediction
CART, Boosting, Random forest, bias-variance tradeoff. Various applications, application to ALS progression based on longitudinal lab data.

Topic 5: Deep Learning
Basic ideas of deep learning, convolutional neural networks, recurrent neural networks, back propagation, various applications in genomics and genetics.

Topic 6: Networks and graphical models
(1) concepts about networks; (2) network models; (3) modeling of the vertex attributes (Markov random _fields; nearest neighbor prediction; (4) modeling of the links (informal scoring; association networks; random graph models); (5) networks clustering and community detection.

Topic 7: Special topics to be determined.