Lecture 1: Introduction

UNIVERSITY OF SAN FRANCISCO

James D. Wilson

MSAN 628

Computational Statistics

Plan for this Lecture

- A little about me
- Overview of Computational Statistics
  - Motivation and Applications
- Overview of this course

A Little About Me

- Ph.D. Statistics and Operations Research (UNC Chapel Hill, ’15)
  - Research focused on statistical analysis of networks
  - Explore, model, and analyze network data (e.g., social networks)
- M.S. Mathematical Sciences (Clemson University, ’10)
- B.S. Mathematics and Chemistry (Campbell University ’08)

A Little About Me

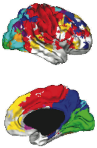
Classes I teach:

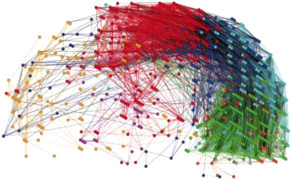
- BSDS 100 - Intro to Data Science with R
- MATH 106 - Business Statistics
- MATH 370 - Probability with Applications
- MATH 373 - Statistical Learning
- MSAN 601 - Linear Regression Analysis
- MSAN 628 - Computational Statistics
- MSAN 700 - Social Network Analysis

A Little About Me

- Born and raised in NC (near Raleigh)
- Live in Rockridge, Berkeley.
- A huge college basketball fan! (Go Heels!)
- Have loved college football since 2008 (Go Tigers!)
- Enjoy tasting beers (bourbon-barrel stouts are my favorite).
- I've always enjoyed teaching..

Research that I work on: Network Analysis





Network Anaylsis Research

- Significance-based community detection
- Generative models for fMRI correlation networks
- Network surveillance and changepoint analysis
- Effects of Networks on testing and inference
- Applications: Silicon Valley Wage Cartel, Urovirulence networks, Student social networks, etc.

A Data Scientist's Toolkit

Harvard's data science toolkit:

1. **Wrangle the data**: gather, clean, and sample data
2. **Manage the data**: access big data quickly and reliably
3. **Explore the data**: to make a hypothesis
4. **Make predictions**: statistical methods
5. **Communicate the results**: visualization, presentations, summaries

What is Computational Statistics?

1. ”... the interface between statistics and computer science" - Wikipedia

2. ”[A field devoted to] the design of algorithms for (1) implementing statistical methods on computers, including the ones unthinkable before the computer age and (2) coping with analytically intractable problems" - Carlo Lauro (IASC)

What is Computational Statistics?

Refers to computationally intensive statistical methods including

- resampling methods
- Markov chain Monte Carlo (MCMC) methods
- local regression
- kernel density estimation
- artificial neural networks
- generalized additive models

A Major Component in Modern Data Science

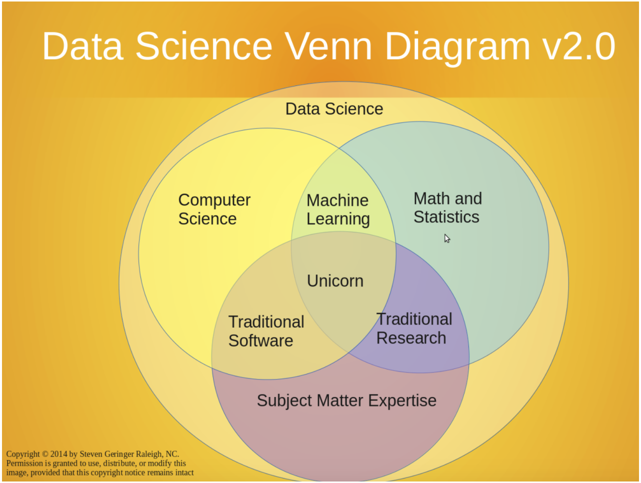


Figure: From *www.datasciencecentral.com*

Sample / Measurement Data

Experiment: Make p measurements on each of n samples.

Result: Data matrix / table X with n rows and p columns

- ith row of X is the vector of measurements on the ith sample
- jth column of X is the vector of values of the jth variable (measurement) across all samples

Different Perspectives on data:

- $n \times p$ matrix X
- n vectors of dimension $p \Leftrightarrow$ samples
- p vectors of dimension $n \Leftrightarrow$ variables

Notation

Data matrix: $$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Rows of X: p variable measurements for each observation.

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$$

Columns of X: n observations of each variable.

$$\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$$

Can write X as: $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = (x_1^T, x_2^T, \ldots, x_n^T)^T$

Data Dimensionality

Old Paradigm: More samples than variables ($n >> p$)

- Number of samples moderate (10s or 100s)
- Number of variables small (1s or 10s)

High Dimensional Paradigm: More variables than samples ($p >> n$)

- Number of samples moderate or large (100s or 1Ks)
- Number of variables *very* large (10Ks or 1Ms)

Big Data Paradigm: Many samples and/or many variables

Source of data: high-throughput measurement technologies for microarray analysis, e-commerce data, click-through rates, etc.

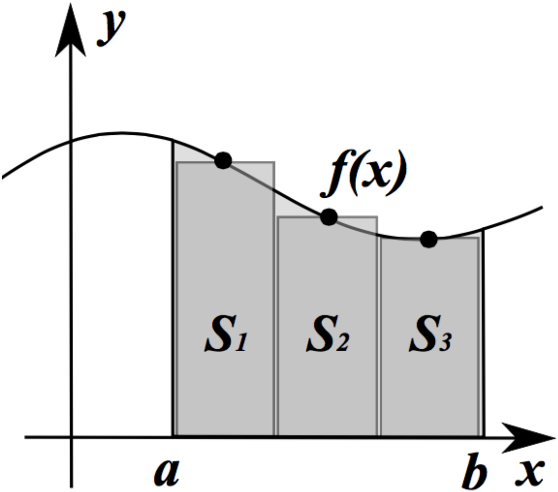
Data Analytic Process

1. Ask what kind of data? Supervised or unsupervised problem? What question are we trying to answer?
2. Prepare / clean data: imputation, outlier removal, etc.
3. Explore data → hypotheses about X and/or model f
4. Apply models and algorithms to answer question
5. Validation of approach

Where Computational Statistics Comes In

1. Imputation of missing data
2. Approximating otherwise intractable functions
3. Simulation
4. Optimization / Estimation

Example: Numerical Integration



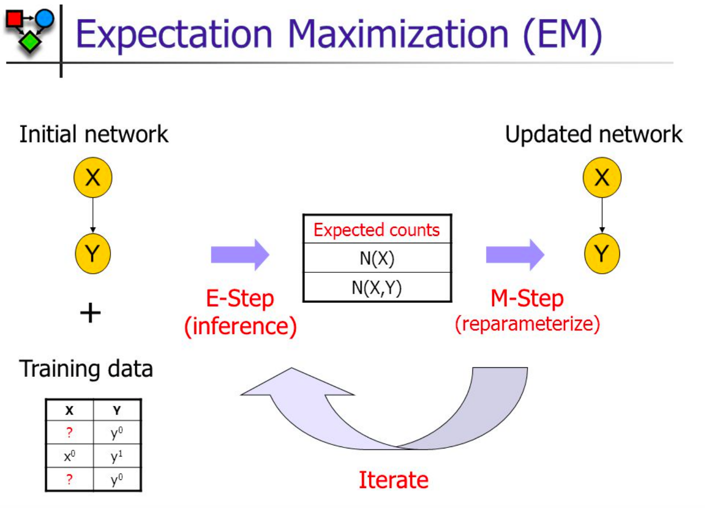
Example: Approximation via Variational Inference

Variational Inference

(in three easy steps…)

1. Choose a family of variational distributions $Q(H)$.
2. Use Kullback-Leibler divergence $\mathrm{KL}(Q||P)$ as a measure of 'distance' between $P(H|D)$ and $Q(H)$.
3. Find Q which minimizes divergence.

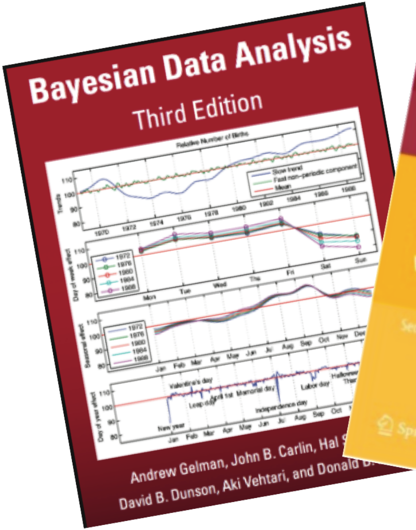
Example: Imputation via the EM Algorithm



Great Resources

- Flowingdata.com
  - Contemporary visualization and data manipulation techniques
- Kaggle.com
  - Kaggle competitions: win money for solving problems!
- Coursera.org
  - Free online courses in data science and machine learning
  - Recent notable course: ”The Data Scientist’s Toolbox”

Primary books in this course



Focus of this course...

Principles of Computational Statistics

- Multivariate probability
- Bayesian computation (multivariate analysis)
- Algorithms: imputation, simulation, estimation, MCMC, EM, variational inference
- Theory: when to use an algorithm and why
- Software: `R`
- Data-driven

...so that we can do more than this...