Introduction to Modern Regression and Predictive Modeling

Merlise Clyde

1/11/2017

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Xin (Julia) Xu
  - Victor Peña
- Course Websites:
  - Main `http://stat.duke.edu/courses/Spring17/sta521`
  - Sakai `https://sakai.duke.edu/portal/site/sta521-s17`
  - Github `https://github.com/STA521-S17`

Grading

Component	Percentage
Participation	10%
Homework	30%
Midterm 1	20%
Midterm 2	20%
Final Data Analysis Project	20%

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory
- Expectations and roles
  - Everyone is expected to contribute equally
  - Everyone is expected to understand *all* code turned in
  - Individual contribution evaluated by peer assessment

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honourably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stackexchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework
  - Group based, everyone is equally responsible
- Late Homework Policy:
  - One day -50%
  - Two or More 0%
- 2 In-Class Midterms

Reproducible Research / Data Analysis

- Unix shell
- R + RStudio + JAGS
- Rmarkdown/knitr
- Git + github

For Friday

- Install recommended software
  - R
  - Rstudio
  - JAGS
- Try R Code School if you are new to R
- Create a github account (if you do not have one already)
- Complete the course survey (email link tonight)

Data Science

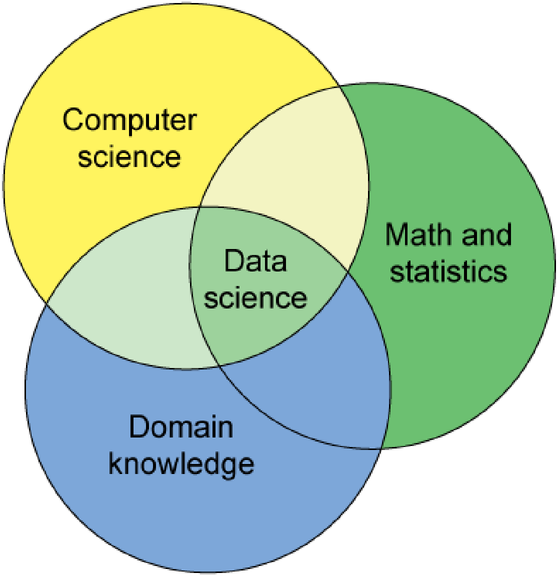

Figure 1

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)
- Goals:
  - learn a model to **predict** Y_i given X_i at new inputs X_i
  - understand relationship between X_i and Y_i (**inference**)

Course Expectations

- Expect to deal with simple to increasingly messy data (real world)
- Writing `R` and `JAGS` code that is reproducible
- self-documented using `Rmarkdown`
- use of version control (git) for team based reproducible coding

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices
- Model Uncertainty (Bayesian Model Averaging and other Ensemble Methods)
- Bayesian Shrinkage and Penalized Likelihood Estimation (Ridge Regression/ LASSO/ Horseshoe)
- Robust Estimation
- Classification and Tree Based Models
- Nonparametric Regression Methods

Themes

- Interpretability versus predictive performance
- Bias-Variance Tradeoff
- In sample versus out-of-sample
- point estimates versus uncertainty quantification
- exact analysis versus appproximation (computational scaling)
- understanding structure of data (relationships)
- Bayesian versus Frequentist ?

Tradeoffs...

All models are wrong, but some may be useful George Box

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
- For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers
- for problems with complex designs and/or missing data Bayesian methods are often better easier to implement (do not need to rely on asymptotics)
- For problems involving hypothesis testing or model selection Frequentist and Bayesian methods can be strikingly different.
- Frequentist methods often faster (particularly with "big data") so great for exploratory analysis and for building a *data-sense*
- Bayesian methods sit on top of Frequentist Likelihood

Important to understand advantages and problems of each perspective!

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
- 17 established SNPS (genetic markers)
- other risk factors (age, family history, oral contraceptive use, number of pregnancies . . .)
- Case - Control design
- variability across study sites (random effects)
- 80% subjects had at least one variable with missing data
- Missing at random versus missing not-at-random
- Focus is on prediction, but still need an interpretable model

EDA, Model Building, and Predictive Checking crucial

Lab Friday

getting started with

- github
- Rstudio
- teams
- data

Complete Survey - Background, Hopes and Expectations!