Introduction(week 1, part 1)

4 Jan 2023

Table of contents

Logistics

- (almost) everything at course web page, https://bbolker.github.io/stat790
- Zoom link: https://tinyurl.com/stat790-zoom-2023
- communication/forums (TBD)
- assignment marks (TBD)

Integrity

- notes on honesty
- why copying code is good
- Stack Overflow, ChatGPT, and all that
- group work

Materials

- Two books:
  - Hastie, Tibshirani, and Friedman (2009) (approx ch. 1-10, 14-15)
  - Shalizi (2022) (approx. ch. 1-4, 7-9, 14-15, 23?)
- Primary literature TBD
- Why both?

Hastie, Trevor, Robert Tibshirani, and J. H Friedman. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York: Springer. http://public.eblib.com/EBLPublic/PublicView.do?ptiID=437866.

Shalizi, Cosma Rohilla. 2022. *Advanced Data Analysis from an Elementary Point of View*. https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

Prerequisites

From the course outline:

- linear algebra (pref. numeric: MATH 3NA)
- probability & statistics (at least up to regression: STATS 3A03)
- computation (pref. R, possibly Python/Julia/Matlab/etc.)
- ML/data science (STATS 780)

See also ADA p. 15.

Goals

- understand theory behind (novel) SL/ML methods
- read papers
- (choose methods)
- implement methods
- read/understand/improve existing methods

Technical skills & tools

These are not focal to the course, but are unavoidable and useful

- reproducibility: version control (Git/GitHub)
- machinery
  - R, Julia
  - VSCode
  - Unix shell?
  - reproducibility: Quarto/Sweave/Jupyter notebooks
- *command line bullshittery* ("bullshit (read: diagnosing and debugging weird things) is a part of life in the world of computers") (Adar 2015)

Adar, Eytan. 2015. "On the Value of Command-Line 'Bullshittery'." *Medium*. https://medium.com/@eytanadar/on-the-value-of-command-line-bullshittery-94dc19ec8c61.

About me

- weird background (physics u/g, Zoology PhD, epidemiological modeling)
- biases/interests:
  - scientific inference $\gg$ pure prediction (but see Navarro (2019))
  - generative models
  - data visualization
  - solving problems in context, practical issues
  - rants (p-values, snooping, dichotomization, unbalanced categorical responses …)

Navarro, Danielle. 2019. "Science and Statistics." Aarhus University. https://slides.com/djnavarro/scienceandstatistics.

Overview of material/themes

what is "statistical learning'' anyway?

- statistics $\to$ computational statistics $\to$ machine learning $\to$ AI
  - inference vs. prediction
  - parametric vs. nonparametric
  - pure analysis vs. pure computation

- small/medium vs. big data

big picture

- make connections between methods
  - analogy: t-test/ANOVA/regression vs "the general linear model" (Lindeløv 2019)
- building blocks/themes

Lindeløv, Jonas Kristoffer. 2019. "Common Statistical Tests Are Linear Models (or: How to Teach Stats)." https://lindeloev.github.io/tests-as-linear/.

basis construction/feature engineering

- SL typically uses **nonparametric** methods
  - families of curves/distributions that can be expanded/made arbitrarily complex
- basis construction
  - splines, Gaussian processes, tree splits, wavelets, Fourier bases, neural network architecture …

optimization

- (stochastic) gradient descent
- iterative methods, e.g.:
  - IRLS (iteratively reweighted least squares), expectation-maximization
  - quasi-Newton (Broyden-Fletcher-Goldfarb-Shanno)
  - expectation-maximization
- map/reduce

bias/variance tradeoff

- under/overfitting, Goldilocks
  - shrinkage, regularization, penalization, dropout (neural networks), tree depth, learning rate (boosting), early stopping, priors …

quantifying accuracy and uncertainty

- raw point estimates are problematic
- accuracy measures
- uncertainty measures
- methods:
  - parametric methods
  - train/validate/test (cross-validation etc.); out-of-bag error
  - bootstrap?
  - high-dimensional inference (Dezeure et al. 2015)
  - conformal prediction (Shafer and Vovk 2008)

sparsity

- dimension reduction
  - structural zeros within vectors or matrices e.g. adjacency matrices
  - reduced-rank matrices
- computational efficiency
- a means to an end

references

Dezeure, Ruben, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen. 2015. "High-Dimensional Inference: Confidence Intervals, p-Values and R Software Hdi." *Statistical Science* 30 (4): 533–58. https://doi.org/10.1214/15-STS527.

Shafer, Glenn, and Vladimir Vovk. 2008. "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research* 9: 371–421. https://jmlr.csail.mit.edu/papers/volume9/shafer08a/shafer08a.pdf.