Applied Topics: Learning from Big Data

Fall 2019 CD; 8:00 A.M. MWF Physics 360 PEngel 319/236

Instructor:

Name: Dr. Tom Kirkman
Office: PEngel 132/6 Phone: 363–3811 email: tkirkman@csbsju.edu
Office Hours: by appointment Drop-by Informal Office Hours: 7:30 A.M. – 5:30 P.M.

Required Texts:

- *An Introduction to Statistical Learning: with Applications in R*
  by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (Springer, 2017)
  Chapters: 1–5, (6), 8, 10
- *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*
  by Paul Teetor (O'Reilly, 2011)
  Chapters: 2, 5, 8, 9, (10), 11, 13, & reference
- `http://www.physics.csbsju.edu/360/`
- `https://github.com/csbsju-physics/360`

Why This Course:

Data Scientist: The Sexiest Job of the 21st Century

Harvard Business Review (2012)

If you use Google Trends to plot the search term "data science" from 2012 to present you'll find an exponential function. For example: `lm(log(y)~x+I(x^2))` results in the polynomial: $0.4027 + 1.1266x - 0.0789x^2$, with highly significant coefficients, adjusted $R^2 = 0.9848$, and AIC$= -78.8$. Clearly data science is hot, but, incidentally, not nearly as hot as "Kim Kardashian" (which is clearly, I hope, a topic of zero lasting importance). If we believe this model, "data science" peaked in $x = 1.1266/(2 \cdot 0.0789) = 7.14$, i.e., February 2019. If we change the model to include cubic terms, "data science" increases without bound. How can we use such models to predict the future if personal choice (of the number of terms) controls future predictions? Perhaps "data science" and "Kim Kardashian" do belong together: more hype than content.

Christmas last provided me with two positive reasons to teach data analytics. First, my sister, who had just finished teaching Math Stat II at wfu.edu, found herself doing the usual (writing letters of recommendation for grad school for her students) but in an unusual context: the letters were not to math or stat graduate programs, rather all were to business analytics programs. Second, at the January meeting of the American Astronomical Society, many talks were devoted to the future of astronomy. While some talks were devoted to new instruments, many were highlighting the need to train astronomers to analyze the huge data sets that are now and will be produced. Just as the Human Genome Project (1990–2003) created a new type of biology (genomics) and a new type of biologist, so new skills are needed now in astronomy.

This course is my first tentative response to these new demands for data analysts. You should be aware of the provisional nature of this course: I won't know what doesn't work until after this course is over! You are my training set; however you are a training set with a voice. I want to encourage you to speak out when things go poorly, so mid-course corrections can be applied. (Incidentally

changing the training while looking at the current outcomes is a terrible practice. Practice what I preach not what I do.)

Topics:

Catalog: Topics covered will vary from year to year. One such topic is physics of solids: crystal structure, lattice vibrations, band theory and electrical conduction in metals and semiconductors. Other topics such as magnetic and dielectric properties as time permits.

Big Data is really about computers: efficient software and fast hardware— topics I'm not going to discuss. (If you want to learn about fast, scalable algorithms for linear algebra, talk to Dr. Mike Heroux, a world-class expert.) Instead we will be using existing software and hardware. In addition to commercial software applications, there are two huge open source solutions: python (particularly scikit-learn) and R. Typically R is the choice of statisticians and python is the choice of astronomers. Oddly enough I'm teaching you R. Most Fridays we will be in PEngel 236, practicing R using the labs in our main textbook (*An Introduction to Statistical Learning* =ISLR). I've assigned the second textbook (*R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics* =RCook) as a general reference for the R language. Frankly the best help source is Google, but often it's necessary to understand (or have examples of) the basic vocabulary before you can compose a successful Google search.

The fundamental ideas of data science are introduced in ISLR chapter 2 as 'trade-offs:' predictive accuracy vs. interpretability (p. 24), predictive accuracy vs. model flexibility (or complexity or degrees of freedom) (p. 32), and overfitting (p. 22). The Bias-Variance Theorem (p. 33) provides a definitive, if perhaps too restrictive, context for these trade-offs. FYI: this 'applied' class will not emphasize proof (often proofs will be relegated to optional videos), rather my aim is that you will learn how to apply statistical methods to practical problems.

Succeeding chapters cover various methods including: regression, K-nearest neighbors, logistic regression, linear and quadratic discriminant analysis, cross validation methods, bootstrapping, brief discussion of regularization methods, dimension reduction with principal component analysis, various tree-based methods including random forests, and clustering methods including K-means and hierarchical. Unfortunately we will not have time for some of the modern (if mathematically complex) methods like support vector machines (chapter 9) and deep neural networks. In fact, given that this is a first-run, 2-credit course, I merely hope that we will have time for all the methods I've listed above! Similarly we will also not be exploring the great visualization abilities of R using `ggplot2` or various discipline-specific packages like bioconductor (for genomics) or the Astrostatistics and Astroinformatics Portal (for astronomy).

Grading:

Your grade will be determined by averaging three scores: total homework score, take-home mid-term exam score, and a final project score (which substitutes for a final exam). Assigned homework is due on the first Monday following assignment. Late homework is highly discouraged (and may be down-graded) but is often accepted. Exam date is: November 15. The final exam period will be used for your individual 10-minute project presentations. The registrar has scheduled the final exam for 8:00 A.M. Thursday, December 12.

Lab:

Most Fridays we will meet in PEngel 236 to practice *R*.

Schedule:

Class		Date	ISLR	Topics	RCook
1	F	Oct 18	1	probability, Bayes, distributions, CLT, contingency tables, hypothesis testing	8
2	M	Oct 21			9
3	W	Oct 23			
4	F	Oct 25	2.3–5	Lab PEngel 236	2,5
5	M	Oct 28	2	accuracy, interpretability, flexibility, overfitting, bias-variance, train/test, regression, classification, K-nearest neighbors (KNN)	11
6	W	Oct 30			
7	F	Nov 1			
8	M	Nov 4	3	regression, outlier, leverage, collinearity, power	11
9	W	Nov 6			
10	F	Nov 8	3.6	Lab PEngel 236	
11	M	Nov 11	4	classification, logistic regression, linear & quadratic discriminant analysis, ROC, KNN	13.7
12	W	Nov 13			
13	F	Nov 15	4.6	Lab PEngel 236	
14	M	Nov 18	5	cross validation, LOOCV, k-fold, bootstrap	
15	W	Nov 20			
16	F	Nov 22	5.3	Lab PEngel 236	
17	M	Nov 25	6.3–4, 10.2	principal component analysis	13.4
18	M	Dec 2	8	trees: decision, regression, classification; bagging, boosting, random forests	
19	W	Dec 4			
20	F	Dec 6	8.3	Lab PEngel 236	
21	M	Dec 9	10.3	clusters; K-means, hierarchical, cutree	13.6
	R	Dec 12		your presentations: 8:00 A.M.	

References:

- STAT 430: Basics of Statistical Learning (uiuc.edu)
  - `https://daviddalpiaz.github.io/stat430fa17/` — class
  - `https://daviddalpiaz.github.io/r4sl/` — online text: *R for Statistical Learning*
- SDS 293: Machine Learning (smith.edu)
  - `http://www.science.smith.edu/~jcrouser/SDS293/lectures/` — lecture slides
- *R for Data Science* by Garrett Grolemund & Hadley Wickham (tidyverse)
  - `https://r4ds.had.co.nz/` — free online textbook
- *Introduction to Data Analysis with R* — IRSA workshop (umn.edu)
  - `https://irsaatumn.github.io/RWorkshop18/`
- *Programming in R for Analytics* — (cmu.edu)
  - `http://www.andrew.cmu.edu/user/achoulde/94842/` — entire course
- *An Introduction to Statistical Learning* — free PDF of our textbook (but buy the hardcopy!)
  - `http://www-bcf.usc.edu/~gareth/ISL/`

Links to Institutional Policies:

- Course Attendance policy
  www.csbsju.edu/academics/catalog/academic-policies-and-regulations/courses/class-attendance
- Statement on accommodations for students with disabilities
  www.csbsju.edu/student-accessibility-services/information-for-faculty/syllabus-statement
- Academic Misconduct and Plagiarism
  www.csbsju.edu/academics/catalog/academic-policies-and-regulations/rights/academic-misconduct
- Sexual Misconduct
  www.csbsju.edu/human-rights/sexual-misconduct/sexual-misconduct-policy
- Title IX policy
  www.csbsju.edu/joint-student-development/title-ix