

STA 521: Predictive Modeling

Duke University, Fall 2015

Instructor: Rebecca Steorts, Assistant Professor, Dept of Statistical Science, beka@stat.duke.edu

Head TA: Abbas Zaidi, PhD Student, Dept of Statistical Science, abbas.zaidi@duke.edu

Office hours Abbas: Thursday, 10-11 am, Old Chem 211A.

Course TA: Yikun (Joey) Zhou, MS Student, Dept of Statistical Science, yikun.zhou@duke.edu,
Office hours: MTuTh: 2 - 3 pm, Old Chem 211A.

Office hours Yikun (Joey): Thursday, MTuTh: 2 - 3 pm, Old Chem 211A.

Course Time: T/Th: 8:30 am - 9:45 am

Lab Time: Monday: 11:45 am – 1:00 pm, taught by Mr. Abbas Zaidi

Course and Lab Location: 071 Perkins

Steorts Office Hours: T/Th: 12:00–1:00

Course webpage: <https://stat.duke.edu/~rcs46/predict.html>

Predictive modeling and applied machine learning methods are an increasingly important tools in both industry and academia. We will start by understanding the basics of data science, what this is and how it's important in the modern data. We then will learn tools that will aide us in predictive modeling and data mining (science) such as reproducible research through Markdown, RStudio, and version control (git). You will be responsible for learning these. Then we will delve into unsupervised and supervised learning, where the data may or may not be normally distributed. The data is often highly dimensional in the covariates or parameter space, calling for dimension reduction. We will explore a range of approaches including factor analysis, principal component analysis, shrinkage methods, and then moving along to data mining techniques such as classification and clustering methods. Time permitting we will cover some special topics.

Prerequisites: Students are expected to be very familiar with R and will be expected to have learned LaTeX by the end of the course. All reports, exams, etc. should be submitted in Markdown Rmd and .pdf format.

Required Texts:

An Introduction to Statistical Learning with Applications in R, Gareth James, Daniela Whitten Trevor Hastie, and Robert Tibshirani, (2013), Springer.

Bayesian Essentials with R, Second Ed., Jean-Michel Marin and Christian Robert (2014). http://www.amazon.com/Bayesian-Essentials-Springer-Texts-Statistics/dp/1461486866/ref=sr_1_1?s=books&ie=UTF8&qid=1440788999&sr=1-1&keywords=christian+robert+and+marin&pebp=1440789000802&perid=12ZCXES5KKFTQ4T4Z4F6

Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Ed., Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). <http://statweb.stanford.edu/~tibs/ElemStatLearn/>

Grading Policy:

Participation and Labs	15%
Homework	35%
Midterm Exam	25%
Final Exam	25%

Topics covered (which are subject to change)

1. Introduction to predictive modeling and data mining
2. Introduction to reproducible research, RStudio, Markdown, and version control.
3. Information retrieval
4. Page rank
5. Regression
6. Modern Regression (Ridge and Lasso)
7. Modern Bayesian Regression Methods
8. Model selection and validation (frequentist and Bayesian)
9. Classification methods (frequentist and Bayesian)
 - Logistic regression
 - LDA and QDA
 - Classification Trees
 - Bagging
 - Boosting
 - Bayesian adaptive regression trees
 - Bayesian information retrieval
 - Falling rules

Course Policies: Homework assignments will be announced in class (along with the due date). It must be turned in at the beginning of the lecture on the due date. Late homework will not be accepted.

All homework's and take home exams *must* be submitted through Markdown and RStudio. The format for homework submission will be announced and explained in class with instructions uploaded to the course webpage. All other formats other than those not explained in class will not be graded. Submissions via email to the TA's or instructor will not be accepted for credit. Information and tutorial about LaTeX can be found at <https://www.tug.org/begin.html>.

There is a Google course discussion page. Please direct questions about homeworks and other matters to that page. Otherwise, you can email the instructors (TAs and professor). Note that we are more likely to respond to the Google questions than to the email, and your classmates may respond too, so that is a good place to start.

Cell phones should be turned off (or set on silent). Laptops are allowed when we are doing applied examples or labs in class, but otherwise should not be out or be used.

Missing class/exams/work: You are responsible for everything from lecture. Do not depend on the course web page for announcements regarding due dates for homework, changes in schedules, etc. Such announcements will be made in class. Homework assignments will be uploaded to the course webpage along with course readings (please check here frequently for updates).

Students who miss graded work due to a scheduled varsity trip, religious holiday or short-term illness should fill out an online NOVAP, religious observance notification or short-term illness notification form respectively. If you are faced with a personal or family emergency or a long-range or chronic health condition that interferes with your ability to attend or complete classes, you should contact your academic dean's office. See more information on policies surrounding these conditions here, and your academic dean can provide more information as well.

Makeup exams must be approved before the time of the exam and will be given only in case of medical or family emergencies (which must be appropriately documented – see above). All work turned in for a grade must be entirely your own. This particularly relates to homework. You are encouraged to talk to each other regarding homework problems or to the instructor/TA, however the write up, solution, and code *must* be entirely your own solution and work.

Academic Honesty: Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. Cheating on exams and quizzes, plagiarism on homework assignments, projects, and code, lying about an illness or absence and other forms of academic dishonesty are a breach of trust with classmates and faculty, violate the Duke Community Standard, and will not be tolerated. Such incidences will result in a 0 grade for all parties involved as well as being reported to the University Judicial Board. Additionally, there may be penalties to your final class grade. Please review Duke's Standards of Conduct. For more information on the Duke honor code (known as Duke Community Standard), please go to <http://integrity.duke.edu/faq/faq1.html>.

Students with Disabilities: Students who require special accommodations in class or during exams should follow the procedures outlined by the Disability Management Program <http://access.duke.edu/students>. Students with disabilities who believe they may need accommodations in this class are encouraged to contact the Student Disability Access Office at (919) 668-1267 as soon as possible to better ensure that such accommodations can be made.

Privacy Policies: Student records are confidential.