Data Mining: Spring 2019

Statistics 36-462/36-662

Basic Course Information

Instructor:

Sivaraman Balakrishnan,
Department of Statistics and Data Science,
Baker Hall 132K,
siva@stat.cmu.edu
Office hours: Tuesday 4:30-5:30pm, BH132K

Teaching Assistants:

1. Kayla Frisoli (Head TA), OH: Monday 2-3pm, PH223B
2. Riccardo Fogliato, OH: Wednesday 9:15-10:15am, PH223B
3. Spencer Koerner, OH: Wednesday 2:30-3:30pm, PH 223B
4. Shengming Luo, OH: Tuesday 9:30-10:30am, PH223B
5. Pratik Patil, OH: Friday 1-2pm, PH223B

Lectures: Tuesday and Thursday, 1:30pm-2:50pm, PH (Porter Hall) 100.

Overview and Objectives

Data mining is the science of discovering structure and making predictions in data sets (typically, large ones).

Data mining spans the fields of statistics and computer science. Since this is a course in statistics, we will adopt a statistical perspective the majority of the course. Data mining also involves a good deal of both applied work (programming, problem solving, data analysis) and theoretical work (learning, understanding, and evaluating methodologies). We will try to maintain a balance between the two.

Upon completing this course, you should be able to tackle new data mining problems, by: (1) selecting the appropriate methods and justifying your choices; (2) implementing these methods programmatically (using, say, the R programming language) and evaluating your results; (3) explaining your results to a researcher outside of statistics or computer science.

Course Outline

Here is an outline of the course material. It is subject to change, depending on time and class interests. The topics divide broadly into the study of supervised and unsupervised problems.

Supervised learning:

- *Linear regression.* Univariate and multivariate linear regression, viewing multivariate regression from simple univariate viewpoint. The assumptions underlying linear regression, and the corresponding optimality properties (best linear unbiased estimate). Weighted linear regression.
- *Regularized regression.* The bias-variance tradeoff. Outperforming linear regression: shrinkage and ridge regression. The importance of variable selection. Best subset selection, forward and backwards stepwise regression, lasso, least angle regression.
- *Model selection and validation.* Training error and optimism. The validation set approach. Leave-one-out cross-validation, K-fold cross-validation. The one standard error rule. The bootstrap.
- *Classification.* Nearest neighbor classification. Linear regression of an indicator vector. Linear discriminant analysis, reduced rank discriminant analysis and Fishers linear discriminant. Logistic regression, and regularized logistic regression.
- *Trees and boosting.* Classification and regression trees. Boostrap sampling and bagging. Boosting, and the connection to regularized regression.

Unsupervised learning:

- *Clustering.* Dissimilarity, K-means clustering, K-medoids clustering, hierarchical clustering, interpreting clustering trees, different linkages. Determining the number of clusters.
- *Dimension reduction.* Principal component analysis, Singular vector decomposition. Directions of maximal variance, or equivalently, approximating a matrix by another matrix with a given (smaller) rank. Intepretation of principal components, usages, limitations. Multidimensional scaling, isomap, local linear embedding.

Logistics

Prerequisities: The only formal prerequisite is 36-401: Modern Regression. I will assume that you are comfortable with basic probability, statistics, linear algebra, and R programming. Specifically, here is a list of topics that you should be more or less familiar with. If you find yourself looking at this list and you dont know a lot of the topics (I dont mean being rusty, I mean you dont know them at all), then come talk to me.

- *Probability.* Event, random variable, indicator variable; probability mass function, probability density function, cumulative distribution function; joint and marginal distributions; conditional probability, Bayes's rule; independence; expectation, variance; binomial, Poisson, Gaussian distributions.
- *Statistics.* Sampling from a population; mean, variance, standard deviation, median, covariance, correlation, and their sample versions; histogram; likelihood, maximum likelihood estimation; point estimates, standard errors, confidence intervals, p-values; linear regression, response and predictor variables, coefficients, residuals.
- *Linear algebra.* Vectors and scalars; components of a vector, geometry of vectors; vector arithmetic: adding vectors, multiplying vectors by scalars, dot product of vectors; coordinate basis, change of basis; matrices, matrix arithmetic: matrix addition, matrix multiplication, matrix inversion, multiplication of matrices and vectors; eigenvalues and eigenvectors of a matrix.
- *R programming.* R arithmetic (scalar, vector, and matrix operations); writing functions; reading in data sets, using and manipulating data structures; installing, loading, and using packages; plotting.

Attendance: Attendance at lectures is highly encouraged. You'll learn more by coming to lectures, paying attention, and asking questions. Plus, it will be more fun.

Office hours: The weekly schedule for office hours is given above. Please make appointments to meet at other times.

Evaluation: Along with the homeworks above, there will be two exams and a final project. The grading breakdown is as follows:

Homeworks	30%
Exam 1	25%
Final Exam	30%
Final project	15%

Exam 1 will be on March 7th. Final grades will be curved.

Textbook: There are three textbooks for this class, all available online.

The main class textbook is *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani. It will be used throughout the semester to supplement the lecture notes. The book is available for free online at `http://www-bcf.usc.edu/~gareth/ISL/`.

Some of the material we will discuss in class is not included in Introduction to Statistical Learning. Two other useful textbooks are:

Elements of Statistical Learning, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer. This book is available free online at `https://web.stanford.edu/~hastie/ElemStatLearn/`

Applied Predictive Modeling by Max Kuhn and Kjell Johnson. Springer. This book is available for free download on SpringerLink as long as you are on the CMU network.

As we move into advanced topics, other references will be given in class.

Canvas, Gradescope and Piazza

The class schedule, lecture notes, homeworks, etc., will be posted on Canvas.

Homeworks will also be submitted and graded on canvas.

There will be approximately one homework assignment every week, generally coming out Thursdays and due the following Wednesday at Midnight. The assignments will be a combination of written exercises and programming exercises.

You are encouraged to type your homeworks, though scans of mathematical sections are

allowed. *You are responsible for making your homeworks readable by the TAs.* Unreadable homeworks may not be graded. Additionally, a small amount may be deducted from homeworks that are difficult to grade due to careless formatting (for example, printing hundreds of pages of nonsense output from R).

In this course, we will be using Gradescope (as a plug-in tool within Canvas) to grade and provide feedback on assignments and exams. This will allow your graders to provide more timely and effective feedback. It also promotes fairer grading practices by facilitating anonymous grading and question-by-question (rather than student-by-student) grading. In addition, Gradescope makes it easy for you to access and review all your (graded) work. During the semester, students will use Gradescope to (a) submit work online, (b) view feedback and scores on graded work, and (c) make a re-grade request within prescribed guidelines. To access Gradescope, simply log on to our course's Canvas site and click on Gradescope in the left-side navigation bar.

A key step in submitting your work to Gradescope is getting a high-quality, digitized scan of your work. There are numerous scanners around campus, and scans can be made with iOS and Android devices. The following website

`https://www.cmu.edu/teaching/gradescope/`

provides more information on using Gradescope, including how to scan assignments via iOS and Android devices, where to find scanners around campus, and how to submit assignments once scanned.

Be sure to take the following important policies and procedures into account whenever you are submitting work to Gradescope:

- If you are writing your assignment by hand (on paper), be sure to use a dark pencil or pen, and write as clearly as possible.
- If you are preparing your assignment electronically, you must use a system (such as LaTeX) that allows you to properly typeset equations.
- Provide a clear separation and demarcation between different exercises on the assignment.
- When you upload your work to Gradescope, be sure to (a) indicate where each question is located within your submission via the click- and-select interface (if you fail to do so, points will be deducted from your score) and (b) after you submit, review each page of your uploaded submission to make sure everything is clear and legible.

- Give yourself some extra time to prepare and submit your assignment online to Gradescope, especially for the first few assignments when you are still getting familiar with it.
- Keep a soft copy of each scanned assignment for your records.
- Regrading requests will be handled through Gradescope; see Exam and Homework Regrading below.
- If you need help with technical issues related to Gradescope, email: canvas-help@andrew.cmu.edu.
- Important grading policy: if the grader cannot read your submission, there is no way to award any points.

Because we often discuss homework in detail at the beginning of lecture, late homeworks will generally not be accepted without prior arrangement.

Your lowest HW score will be dropped, but see below.

Piazza: Please post any questions regarding homework or the lecture to Piazza. We will do our best to be helpful and respond promptly. *However,* you should be reasonable about expectations of responses to questions on the due dates of homeworks.

Bonus: There are two ways to earn extra credit in this class. If you turn in all HWs and receive a score of more than 60% you will be awarded a small bonus. Similarly, to encourage student participation on Piazza, you will receive a small bonus for actively answering questions on Piazza.

Plagiarism and collaboration:

You are not allowed to utilize any materials from past versions of this course. Any such use will be considered cheating, and treated as such.

Collaboration is a tricky issue. The support and assistance of your classmates can go a long way towards helping you to understand the material. Ultimately, however, you are responsible for preparing yourself for the final exam, later courses, and your future. I encourage you to collaborate on understanding the material and your homework. However, everything you submit must be your own work. Your final writeup should be done independently and you must write your own code for computational problems. You are responsible for understanding everything that you submit. **Please ask me if you have any confusion.** Instances where students copy the work of another student will be treated as cheating.

Accommodations for Students with Disabilities:

If you have a disability and are registered with the Office of Disability Resources, I encourage you to use their online system to notify me of your accommodations and discuss your needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Health & Well-being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night:

1. CaPS: 412-268-2922
2. Re:solve Crisis Network: 888-796-8226
3. If the situation is life threatening, call the police. On campus: CMU Police: 412-268-2323. Off campus: 911