Data Mining: Spring 2013

Statistics 36-462/36-662

Instructor: Ryan Tibshirani, Dept. of Statistics, Baker Hall 229B, `ryantibs@cmu.edu`

Teaching assistants:
Li Liu, `lliu1@andrew.cmu.edu`
Cong Lu, `congl@andrew.cmu.edu`
Jack Rae, `jwr@andrew.cmu.edu`
Michael Vespe, `mvespe@andrew.cmu.edu`

Lectures: Tuesdays and Thursdays 1:30-2:50pm, Porter Hall 125C

Overview and objectives

Data mining is the science of discovering structure and making predictions in data sets (typically, large ones). Applications of data mining are happening all around you—and if they are done well, they may sometimes even go unnoticed. How does Google web search work? How does Shazam recognize a song playing in the background? How does Netflix recommend movies to each of its users? How could we predict whether or not a person will develop breast cancer based on genetic information? How could we search for possible subgroups among breast cancer patients, suggesting different variants of the disease? An expert's answer to any one of these questions may very well contain enough material to fill its own course, but basic answers stem from the principles of data mining.

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.

Outline of material

Here is an outline of the course material. It is subject to change, depending on time and class interests.

Unsupervised problems.

- *Information retrieval and PageRank.* Finding documents relevant to a given query. Representation by bag-of-words, measures of similarity (or distances), searching by similarity, evaluating performances, incorporating user feedback. Applying this to the web, and exploiting link structure via the PageRank algorithm.

- *Clustering.* Dissimilarity and scatter. K-means clustering, K-medoids clustering. Hierarchical clustering, interpreting clustering trees, different linkages, top-down and bottom-up. Determining the number of clusters.

- *Dimension reduction.* Principal component analysis. 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.

- *Correlation analysis.* Correlation. Canonical correlation analysis. Zero correlation versus independence. Shortcomings of correlation for nonlinear relationships. Rank correlation, maximal correlation, distance correlation.

Supervised problems.

- *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) and inferential properties. 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 Fisher's 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.

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 don't know a lot of the topics (I don't mean being rusty, I mean you don't 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.

Class website: The class website is `http://www.stat.cmu.edu/~ryantibs/datamining`. The class schedule, lecture notes, homeworks, etc., will be posted there.

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 below. Please make appointments to meet at other times.
RT: Tuesdays 3-4pm, Baker 229B
LL: Wednesdays 5-6pm, FMS 320
CL: Fridays 11am-12pm, Wean 8110
JR: Wednesdays 11am-12pm, Wean 8110
MV: Mondays 5-6pm, FMS 320

Textbook: The 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. This book has not yet been published and the authors have been kind enough to let us use it, so you must not distribute any of its contents. The book and a corresponding R package are available online at `http://www-bcf.usc.edu/~gareth/ISL`. This page is password protected, use the login name: `StatLearn` and password: `book`.

Another optional textbook, which is similar to the class textbook above but is aimed at a more advanced audience, is *The Elements of Statistical Learning* (Second Edition) by Hastie, Tibshirani, and Friedman. This book contains some material that is not covered by our class textbook and so it could serve as a useful supplement for some of the lectures. It is freely available online at `http://www-stat.stanford.edu/ElemStatLearn`.

Evaluation: There will be 6 homework assignments, approximately one every two weeks. The assignments will be a combination of written exercises and programming exercises. The assignments will be posted on the course website, and your homeworks will be collected at the start of lectures. Your lowest homework score will be dropped.

There will be two in-class midterms. There will be a final project—details on the project to come. The grading breakdown is as follows:

Homeworks	45%
Midterm 1	15%
Midterm 2	15%
Final project	25%

Plagiarism: You are encouraged to discuss homework assignments with each other. But you must submit your own original work, both written work and computer code. Explicitly sharing your written work or code with someone else is not allowed. See the student handbook's section on "Cheating and Plagiarism" (`http://www.cmu.edu/policies/documents/Cheating.html`).