36-350: Data Mining

Fall 2008

Instructor:
Cosma Shalizi, Statistics Dept., Baker Hall 229C, `cshalizi@stat.cmu.edu`

Teaching Assistant:
Joey Richards, `jwrichar@stat.cmu.edu`

Lectures: Monday and Wednesday, 10:30–11:20, Porter Hall 226B
Friday, 10:30–11:20, Doherty Hall 1217

Overview and Objectives

Data mining is the art of extracting useful patterns from large bodies of data; finding seams of actionable knowledge in the raw ore of information. The rapid growth of computerized data, and the computer power available to analyze it, creates great opportunities for data mining in business, medicine, science, government, etc. The aim of this course is to help you take advantage of these opportunities in a responsible way.

This course should give you a thorough introduction to modern data mining. Faced with a new problem, you should be able to (1) select appropriate methods, and justify their choice, (2) use and program statistical software to implement them, and (3) critically evaluate the results and communicate them to colleagues in business, science, etc.

Data mining is related to statistics and to machine learning, but has its own aims and scope. Statistics is a mathematical science, studying how reliable inferences can be drawn from imperfect data. Machine learning is a branch of engineering, developing a technology of automated induction. We will freely use tools from statistics and from machine learning, but we will use them as tools, not things to study in their own right. We will do a lot of calculations, but will not prove many theorems, and we will do even more experiments than calculations.

Contents

This is a rough outline of the material. The first five items should take us to the middle of the semester, and (6) and (7) will definitely be covered; the others will depend on time and class interests.

1. *Searching by similarity*: Searching by content (texts, images, genes, . . .); attributes, representations and definitions of similarity and distance; choice of representation; multi-dimensional scaling; classifications; image search and invariants; user feedback; evaluating searches

2. *Information*: information and uncertainty; classes and attributes; interactions among attributes; relative distributions

3. *Clustering*: supervised and unsupervised learning; categorization; unsupervised category-learning, a.k.a. clustering; k-means clustering; hierarchical clustering; geometry of clusters; what makes a good cluster?

4. *Data-reduction and feature-enhancement*: Standardizing data; using principal components to eliminate attributes; using factor analysis to eliminate attributes; limits and pitfalls of PCA and factor analysis; nonlinear dimensionality reduction

5. *Regression* Review of linear regression; transformations to linearity; the truth about linear regression; local linear regression; polynomial regression; kernel regression; other non-parametric methods

6. *Prediction*: Evaluating predictive models; over-fitting and capacity control; regression trees; classification trees; neural networks; combining predictive models; forests; how to gamble if you must

7. *Classification*: Supervised categorization; linear classifiers; logistic regression; the kernel trick; base rates and multiple testing; spam, fraud, credit cards, profiling

8. *Change over time*: trends and time series models; Markov models; hidden Markov models; adapting to change and failure to adapt

9. *Modeling interventions*: Estimating causal impacts without experiments; matching; graphical causal models and Tetrad.

10. *Waste and Abuse*: what happens when the data are bad; what happens with the wrong kinds of data; situations when data mining will fail; trying not to be evil; some failures

Practicalities

Class We will have lectures on Mondays, Wednesdays and Fridays. (The on-line catalog thinks the Friday class is a lab; it is wrong.) You are responsible for everything in the lectures, even if it is not covered in the assigned readings. I will not take roll; but attending class, paying attention, and participating will help you learn.

Textbook The **required** textbook is *Principles of Data Mining* by Hand, Mannila and Smyth (MIT Press, 2001, ISBN 978-0-262-08290-7). The campus bookstore should have it as of 26 August. Some additional readings will be posted on the class website.

Homework There will typically be one homework set a week, due on Fridays at the start of lecture, either in class or submitted in e-mail. Solutions will be posted to the class website, generally after graded assignments are handed back.

The main point of the homework is to help you understand the material. It is also supposed to encourage you to keep up with the class. Assignments will contain a mixture of calculations, writing questions, and computer exercises. Computer code, or computer output, must always be accompanied by a written explanation (in English!) of what the code does and what the output shows; do not assume that anything meant for the machine is self-explanatory.

Each homework will be worth 100 points, divided roughly evenly over the problems. Your lowest homework grade will be dropped, unless you skip the last assignment, in which case everything will count. (Do not skip the last assignment.)

Late homework may get partial credit at my discretion. Homework turned in after solutions are posted will get no credit. (Please turn in your homework on time.) There can be extensions for the usual reasons given in the student handbook (medical, religious, university event, etc.); please make sure to get the proper forms (as described in the handbook). In special circumstances, please see me *as soon as possible.* (Please turn in your homework on time!)

Computing: R *Every* assignment will contain at least one computer exercise, which will use a software package called R (`http://www.r-project.org/`). R is free, standard for statistical programming, and can run on almost any computer system. If you do not have reliable access to a computer which can run R, let me know as soon as possible.

Some good resources on learning about R include:

- The official intro, "An Introduction to R", available online at `http://cran.r-project.org/doc/manuals/R-intro.html` and `http://cran.r-project.org/doc/manuals/R-intro.pdf`
- John Verzani, "simpleR", available at `http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`

- “Quick-R”, `http://www.statmethods.net/`. This is primarily aimed at those who already know a commercial statistics package like SAS, SPSS or Stata, but others may find it useful as well.
- W. John Braun and Duncan J. Murdoch, *A First Course in Statistical Programming with R* (Cambridge University Press, 2007, ISBN 978-0-521-69424-7)

Exams There will be a mid-term exam and a final exam. Both will be cumulative. You will not need a computer, though you will probably have to interpret R-style computer output.

Office Hours There will be two regular office hours each week: the professor’s (Thursdays, 4 pm, 229C Baker Hall) and the TA’s (TBD). Please make appointments to meet at other times.

Your grade will be 50% homework, 20% mid-term, and 30% final exam.

Plagiarism You’re free, and encouraged, to talk about assignments with each other. But all work, including computer code, that you turn in must be your own. Sharing code or results will result in zero credit and a letter to your dean at the very least. See the student handbook’s section on “Cheating and Plagiarism” (`http://www.cmu.edu/policies/documents/Cheating.html`).