Data Mining
Tools for Exploring Big Data

Robert Stine
Department of Statistics
Wharton School, University of Pennsylvania
www-stat.wharton.upenn.edu/~stine

Modern data mining combines familiar and novel statistical methods to identify reproducible patterns in big data. The objective outcome is prediction. If your model predicts new data better than alternatives, then you've made a contribution. Rather than build a model that relates one or two experimental results to a response, data mining involves searching for patterns. Such searches commonly scan thousands of features, looking for the few that are predictive of the response. The search might be entirely automated or allow expert insight. Once considered vile, data mining has become respectable, useful, and necessary. Data mining is needed when dealing with wide data tables, those having more variables than cases.

This course introduces data mining through a combination of lectures and examples. You'll see examples that look for patterns in voting behavior, patients at risk of a disease, prospective job candidates, and credit applications that reveal fraud. In each illustration, the goal is prediction. Rather than interpret a pattern found in one set of data, the objective is to predict new data. Interpretation is fun, but we'll exercise considerable restraint to avoid confusing association with causation. Even if you stick to simple models, concepts from data mining can diagnose if you've missed an important feature of your data.

Data mining does not require exotic hardware or software. Today's PC would have been a supercomputer in 1999. You can explore large datasets quite well with nothing more than regression and a laptop. Once we develop the fundamentals, you'll be able to appreciate the strengths and weaknesses of exotic methods. The course begins with regression, then covers logistic regression, neural networks, and classification and regression trees, with a bit of cluster analysis.

You need to do data mining to learn data mining. For this class, we'll use a combination of R and JMP from SAS. JMP interactively handles large data sets and includes an extensive collection of algorithms for building and assessing models. The software is highly interactive with amazing interactive graphics. Two class sessions are devoted to lab time so that you can try some of these tools yourself in a supervised lab.

Participants who attend these lectures are encouraged to work on a project associated with an ICPSR data set. The work on that project will help you learn how the tools work. Possible topics for this project include data from studies related to elections, health outcomes, criminal sentencing, economics, and social activities.

Guide to Planned Lectures

Lecture 1. Introduction, Exploring Large Data Sets

Good data analysis begins by looking at the data. That can be hard to do when the dataset has hundreds of columns, but it's not impossible. This lecture introduces data mining and explores several datasets using *interactive graphics*, a key strength of JMP. The objectives are to gain familiarity with the data, spot unusual patterns, recognize collinear variables, and form conjectures. Hypothesis *generation*, not just testing, is an important aspect of data mining.

Predictive models aim to predict or classify new observations. Regression analysis is the most commonly used predictive methodology in statistics and the benchmark for data mining as well. Much of the success that you can have with a regression model comes from the fact that regression comes with well-developed diagnostics combined with tools for inference that can be enhanced with *bootstrap resampling*.

A key challenge in data mining is finding the right set of explanatory variables. We may start from those that come in the data directly, but we usually need to expand this set. Even though we may start with a wide data set, the columns in the data may not be the right ones to use in a data mining analysis. The simplest extensions are interactions, products of the original explanatory variables. Graphical tools known as *profilers* help us explore the structure of models, helping convey the impact of interactions and nonlinearity. Before we can exploit interactions in data mining, we need to be acquainted with what they are and what they do in a model.

Calibration is an important additional diagnostic of the performance of a model, one that is easy to check and useful in applications. In many cases, such as when building models for a 0/1 response, it is far easier to calibrate the original regression model rather than switch to a more elaborate method. This lecture will also introduce some of the data sets used in the class, particularly data from the 2008 US presidential election.

Lecture 2. Data Mining with Regression

Wide data sets challenge the thoughtful modeler. Even the best researcher with a clear theory will wonder if other variables might be predictive of the response. Since most collections of possible explanatory variables are highly correlated, it can be very hard (and risky) to impose a strong interpretation on the estimates. That's where methods that automatically *search for predictors* (explanatory variables that don't really explain but do predict the response) become necessary. The best known, most disparaged, and yet perhaps most useful technique among these is the greedy search provided by *stepwise regression*. Even if you have your heart set on one of the new algorithms, you should always set a baseline for comparison with stepwise regression.

Most datasets contain *missing values*, and this class discusses simple methods for handling these anomalies – that are far too common – when data mining. We'll also borrow ideas from *multivariate analysis* (*e.g.*, principal components and cluster analysis) to find interesting views of data.

Lecture 3. Model Validation

Automated searches for patterns increase the risk of *over-fitting*, adding variables to a predictive model that are not in fact predictive. Over-fitting is common; we've all seen applications in which a model fits the observed sample well but predicts new data poorly. This lecture considers how to control stepwise regression. *Cross-validation* and the roughly equivalent criterion known

as *AIC* are often recommended, but we'll come down on the side of methods related to the *Bonferroni* criterion.

Cross-validation is often used more generally to pick a model. An important example of this use of cross-validation comes in picking the best model offered by the *lasso*, a different way of picking the variables for a model that shrinks the model estimates. Though handy, cross-validation has problems, too. We'll identify situations when it works and others that lead to problems.

Lecture 4. Lab Session

Hands on time in the Michigan Lab.

Lecture 5. Holiday!!!

July 4th. Go to the parade in downtown Ann Arbor.

Lecture 6. Logistic Regression and Classification

Many modeling tasks seek to predict qualitative behavior, such as whether a voter participates in an election or a customer makes a purchase. Models for choices – *classifiers* – require different methods that adjust for the categorical nature of the response. The canonical example is *logistic regression* for a binary response. *Calibration* remains important, and a calibrated linear regression often matches well to a logistic regression. That's good, because searching for the best logistic regression can be very slow!

Changing the response from numerical to categorical also changes the way that we measure how well a model performs. Least squares may be reasonable, but other metrics such as *ROC curves* and *confusion matrices* are popular.

Lecture 7. Neural Networks and Boosting

Regression models produce an equation that "explains" how the model predicts new cases. Alternative methods such as *neural networks* work differently. Neural networks, and the closely related method known as projection pursuit regression, blend several regression models together using heuristics borrowed from engineering and biology. Model visualization – profiling – becomes necessary for understanding what a network does.

Boosting is a general-purpose technique for improving – boosting – the performance of a model. *Boosted networks* are an example of *model averaging* in which you combine several different models to arrive at a prediction (another is known as *bagging*). We'll explore boosting in the context of neural networks, but you should recognize that it can be used effectively in many other situations.

Lecture 8. Classification and Regression Trees, Random Forests

Regression trees (*a.k.a.*, CART with the closely related *classification trees*) are different enough from regression models that they deserve more time to appreciate. In place of an equation, a tree produces a set of rules that determine how to predict or classify the response. This lecture uses trees as alternatives to regression models, considering what they are very good at (*e.g.*, finding interactions and producing a set of rules that is often more appealing than the concept of an

equation) and what they don't do so well (*e.g.*, smooth patterns). Part of the appeal of using trees is the elegant implementation in JMP.

Trees are neat, but if you like them, you'll want to learn how to grow forests. *Random forests* are another example of *mode averaging.* Random forests use *bootstrap resampling* to create multiple data sets that can be used to fit alternative models. Averaging the predictions of these models usually produces a better prediction than any one model alone. The ideas apply widely beyond trees to virtually any predictive model.

Lecture 9. Lab Session

Hands on time in the Michigan Lab.

Lecture 10. Vector Space Models for Text

We'll use this last class to review prior topics and handle the inevitable overflow from prior lectures. We'll also discuss an ongoing project that applies regression to *text mining*: modeling text rather than numbers. Once you see how that's done, you'll see how to expand the scope of models to images and other types of novel data.

Time permitting, we will look at related methods for very large data sets based on ideas known as *kernels* and *random projection*. These ideas have produced a wave of techniques related to principal component regression that can be used when dealing with millions of variables.

References

If you would like to do some reading to accompany these lectures, or perhaps follow up afterwards, here are a few papers, books, and web sites that you might find useful.

Beck, King and Zeng (2000), Improving quantitative studies of international conflict, *American Political Science Review*.

The authors use a neural network to explore improvements in modeling international data. See the follow-up by DeMarchi et al (2004).

Berk, R A (2008), *Statistical Learning from a Regression Perspective*, Springer.

Berk starts from the regression point of view and includes trees and ensemble methods like bagging and random forests. There's also a bit on smoothing via regression.

Berk, R A (2006), An introduction to ensemble methods for data analysis, *Sociological Methods & Research*, **34**, 263–295.

This paper describes ensemble methods like bagging using the example of random forests. There are examples with cross-validation as well.

Breiman, L (2001), Statistical modeling: the two cultures, *Statistical Science*, **16**, 199-215.

Statistics missed the boat, sticking to asymptotic estimates for small samples as the world of computing and large data bases exploded. Several good discussions accompany this article.

Breiman, L, J Friedman, R Olshen, and C J Stone (1984), *Classification and Regression Trees,* Wadsworth.

This classic popularized the use of tree-based models and the use of cross-validation in picking good models. Still a good read.

Chatfield, C (1995), Model uncertainty, data mining, and statistical inference, *Journal of the Royal Statistical Society, Series A*, **158**, 419-466.

Over-fitting is a serious problem when the data suggest the model, often leading to wildly optimistic promises of prediction accuracy.

DeMarchi, Gelpi, and Grynaviski (2004), Untangling neural nets, *American Political Science Review.*

A rebuttal to Beck et al (2000). Maybe you don't need a neural network if you are able to build more predictive, interpretative explanatory variables.

Foster, D P and R A Stine (2004), Variable selection in data mining: building a predictive model for bankruptcy, *Journal of the American Statistical Association*, **99**, 303-313.

Using 3,000,000 months of credit card activity and 67,000 possible features, we show how to use least squares regression to build a model that predicts as well (or better) than the computational learning algorithm known as C4.5. The algorithm is now implemented in the SAS enterprise miner software package.

D P Foster and R A Stine (2008), α-investing: a procedure for sequential control of expected false discoveries, *Journal of the Royal Statistical Society* B, **70**, 429–444.

This paper shows how to use a theory to guide the allocation of the chance for false positives over a sequence of tests, as in the case of stepwise regression.

Friedman, J (2001), The role of statistics in the data revolution, *International Statistics Review*, **69**, 5-10.

Jerry Friedman is one of the most creative modelers in statistics. Here's his take on how statistics can play a more central role in a world of large data sets.

Friedman, J and B Silverman (1989), Flexible parsimonious smoothing and additive modeling, *Technometrics*, **31**, 3.

This paper shows that nonparametric regression can be viewed as a special case of stepwise regression. It's all in the choice of the *x* variables that the search considers.

Hand, D J, H Mannila, and P Smyth (2001), *Principles of Data Mining*, MIT Press.

There are a lot of books on data mining, but most talk more about assembling the data and are written for computer science. Assembling the data is a hard part of the problem, and if you get it wrong, it does not matter how you do the modeling. This book describes some of those issues, but goes on to summarize the statistical methods as well. A nice high-level view of models and patterns in general.

Hand, D J, G Blunt, MG Kelly, and N M Adams (2000), Data mining for fun and profit, *Statistical Science*, **15**, 111-131.

If my predictions are more accurate than yours, I profit and you lose. It's no wonder that data mining has become important in the business world. Be they pharmaceuticals like Merck and Pfizer or web sellers like Amazon, a large chunk of the value of many firms lies in their proprietary data.

Hastie, T, R Tibshirani, and J Friedman (2001), *The Elements of Statistical Learning,* Springer.

Much of the initial development of methods for handling large data sets began outside of statistics, in an area known as computational learning or, more boldly, knowledge discovery. Computer science was quick to see the importance of getting value from large databases, and forged ahead. They certainly came up with better names (e.g., neural network versus projection pursuit regression). You can download this book on-line for free.

Hopkins and King (2010) A Method of Automated Nonparametric Content Analysis for Social Science, American Journal of Polical

James, G, D Witten, T Hastie, R Tibshirani (2013), *An Introduction to Statistical Learning*, Wiley.

This text provides a well-written, engaging introduction to the methods of data mining (with the exception of trees). Basically, it's a simplified, more introductory (ie, less superfluous mathematics) version of *Elements of Statistical Learning*. The discussion of each topic comes with detail examples using R to analyze data. If there's a weakness, its the data analyzed in these examples has a lot more structure than what you often find in survey responses. Its nice to see, however, how things work in a good situation.

Ledolter, J. (2013), *Data Mining and Business Analytics with R*, Wiley.

Wide range of topics from association rules, cluster analysis, decision trees and various types of parametric regression models. Chapters are short with a brief overview of the relevant theory, then puts the emphasis on examples with R commands and output. Includes a chapter on a particular method of text mining.

Linoff, G S, and M J A Berry (2011), *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management* (3rd Edition), Wiley.

Emphasizes role in business decision making, with 100+ pages introducing the role of data mining. Wide scope of methods with discussion of presentation, communication of results, and privacy issues.

Monroe, B L, M P Colaresi, and K M Quinn (2008) Fightin' words: lexical feature selection and evaluation for identifying the content of political conflict, *Political Analysis*, 16, 372–403.

Uses the frequency of specific words to discover political affiliation of speaker from text.

MacKay, D J C (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

This text introduces many of the principles that guide modern research on modeling within computer science, nicely called "machine learning." The text introduces the foundations by developing the connection between statistics and information theory. These connections provide heuristics for the computing algorithms.

Sala-i-martin (1997), I just ran two million regressions. American Economic Review.

Not particularly germane to our treatment of data mining, but I love the title.

Stine, R A (1989), An introduction to bootstrap methods: examples and ideas.

Introduces fundamental ideas, heuristics of bootstrap resampling.

Stine, R A (2004), Model selection using information theory and the MDL principle, *Sociological Methods & Research,* **33**, 230-260.

You have to think about the scope of the search when you build a model by finding the best fit possible using anything from a wide database. Information theory (ideas that show how to fit more data on your computer disk or make cellular telephones work) turns out to offer a very useful paradigm for judging models.

Tan, P N, M Steinback, and V Kumar (2006), *Introduction to Data Mining*, Pearson.

This is a wide-reaching, medium-level math introduction. The presentation is not nearly so technical as Hastie. It includes many of the ideas that are now used in computer science for modeling, such as the so-called kernel trick and support vector machines. (It has little discussion of regression or logistic regression.)

Torgo, L (2011), *Data Mining with R*, CRC Press.

Introduces data mining and R five case studies, but some of the data is really small (like $n = 7$). Examples include nature, stock trading, fraud detection, and microarrays. The methods include the usual ones (like regression), but reach out to support vector machines (SVM) and multivariate adaptive regression splines (MARS).

Turney, P D and P Pantel (2010), From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141-188.

Excellent introduction to VSMs as used in text analysis, offering a linguistic perspective on how simple counts can possibly reveal the meaning of speech.

Ward, Greenhil, and Bakke (2010), Perils of policy by p-value: Predicting civil conflicts *Journal of of Peace Research*.

Addresses the problems introduced by overfitting, particularly treating statistical significance too literally when interpreting the results of fitting many models.

Witten, I H, E Frank, and M A Hall (2011), *Data Mining: Practical Machine Learning Tools and Techniques* (3rd Edition), Morgan Kaufman.

The emphasis here is on trees and association rules. The coverage is terse, but comprehensive with much of the terminology mentioned. Technical material is included, but confined to boxes. The examples are closely tied to the WEKA software tools. Examples use small data sets.

www.kdnuggets.com

This web site links to software, data and conferences, including the data sets used for its annual competitions to see what sort of data mining software can predict a hold-back sample the best.

Witten, I H, E Frank, and M A Hall (2011), *Data Mining: Practical Machine Learning Tools and Techniques* (3rd Edition), Morgan Kaufman.