Data Mining
ICPSR Summer Program, 2008

Robert Stine
Statistics Department
Wharton School, University of Pennsylvania
Philadelphia, PA 19104-6340
stine@wharton.upenn.edu
www-stat.wharton.upenn.edu/~stine

These lectures introduce data mining. Once a nasty thing to be accused of, data mining has become respectable, useful, and even necessary. What is data mining? Basically, data mining refers to statistical algorithms that seek reproducible patterns in "wide" data sets. Wide data sets have many columns (or variables), perhaps even more columns than rows. The standard for reproducibility is prediction. If you can predict new data accurately, then you've found a reproducible pattern. Rather than building a model that relates one or two carefully chosen measurements to a response, data mining usually involves a search for patterns from a wide dataset. Such searches might consider the use of 100,000 or more features, looking for the few that are predictive of the response.

In these lectures, you'll see examples that use data mining to identify patients who are most at risk of a disease, prospective job candidates who are most likely to succeed in their career, and credit applications that suggest fraud. In each of these illustrations, the goal is prediction. Rather than try to interpret a pattern found in one set of data, the objective is to predict results in new data. Sure, we'll talk about the models – and look at them as well – but we'll use considerable restraint to avoid confusing association with causation.

Data mining does not require exotic hardware or software. You can mine data quite well with least squares regression and a laptop. Indeed, once you understand how to use regression for data mining, you'll be able to appreciate the strengths and weaknesses of other methods. So, I'll start with regression, and then move on to logistic regression, classification and regression trees, and a bit of neural networks and cluster analysis. I'll use each of these during the lectures.

For demonstrations in class, I'll use JMP (from SAS) and R. JMP costs about $65, handles very large data sets, and – depending on the version – includes a nice collection of regression tools along with data mining tools such as trees and neural networks. The JMP software is downloadable via the web at www.jmp.com. R is covered in other summer program courses and is available free.

Lecture guide

Lecture 1. Introduction

Data mining uses some type of automated search to identify a predictive model from the features in your dataset. The problem with this approach is that models fit the data that suggest their form better than they fit new observations, even if all of your data come from the same population. If you have ever used stepwise regression, you have probably run into this problem. Your model looked great on the data that you used to build it (the training data), but predicted a hold-back sample poorly (the validation data).

To be successful, the data miner has to avoid this *overfitting*. Just as it can be difficult to separate cause from coincidence, it can be hard to recognize overfitting.

Fortunately, there are approaches that avoid overfitting that are simple to employ. A very effective way to avoid overfitting uses a Bonferroni *p*-value rather than the familiar 0.05 criterion.

Lecture 2. Data mining with regression and logistic regression

This lecture introduces what I will call the 4 C's of data mining. Once you have built your model, you need to consider these properties before you go any further. In July, we might decide that we need a few more!

C^1 Cost

Does the model achieve the accuracy that you need? How well does it classify observations, if that's the goal? How do you convert predictions to decisions?

C^2 Calibration

Are the predictions calibrated? For the days your model predicts a 40% chance of "rain", does it rain on 40% of them? Or, like most weather forecasters, is your model not calibrated?

C^3 Comparison

Would a simpler model work just as well? You should always have a baseline model for comparison.

C^4 Cross-validation

Unless you try to predict new data, it can be very, very hard to know just how well your model will perform when applied to new cases. Better to find out sooner rather than later.

Lecture 3. Classification and regression trees

Regression and related methodologies produce an equation. CART produces a tree, a sequence of "questions" that group similar cases. As models, trees have certain advantages, such as the ability to find interactions and the ease of explanation. They also have a few weaknesses compared to regression models as well.

Lecture 4. Growing trees and recent advances

We'll spend more time working with trees and see how trees complement regression models. Research has not stopped with trees. We'll also take a peek at neural networks and cluster analysis and discuss methods like model averaging and random forests that combine the predictions of many models.

Some references for background and further reading

If you would like to do some reading before coming to these lectures, or perhaps afterwards, here are a couple of papers, books, and web sites that you might find useful.

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

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

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

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.

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

Foster, DP and RA 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 has been 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 Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, **31**, 3-a lot.

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, DJ, H Mannila, and P Smyth (2001). *Principles of Data Mining*. MIT Press, Cambridge MA.

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, DJ, G Blunt, MG Kelly, and NM 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, New York.

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).

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

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.

Stine, RA (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, PN, M Steinback, and V Kumar (2006). *Introduction to Data Mining*. Pearson, Boston.

This is a wide-reaching, medium level introduction to the area. The presentation has some math and covers a lot of the area, but this book is not nearly so technical as Hastie. 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 very little discussion of regression or logistic regression.)

www.kdnuggets.com

This web site provides a range of 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.