

Lecture 1: What is Data Mining?

Alan Lee

Department of Statistics
STATS 784 Lecture 1

July 26, 2017

Outline

Housekeeping

Introduction

Data mining projects

Statistical learning

References

Plan of today's lecture

In today's lecture we will discuss the course organization and then cover some introductory material on data mining.

Taught by

- ▶ Alan Lee: First half of semester
- ▶ Thomas Yee: Second half

My contact details

Office: Rm 367, building 303S

Phone: Extn 88749, DD 09-923-8749

Email: lee@stat.auckland.ac.nz

Course aims

- ▶ To introduce you to the data mining process
- ▶ To teach you some common (and useful!) data mining tools
- ▶ To give you some data wrangling skills

Course details

- ▶ Two one-hour lectures per week, Monday at 4pm, Rm 508 in building 201E (Human Sciences) and Friday at 11am Rm G16 in building 303 (Maths/Stats)
- ▶ One one-hour in-class lab, Monday at 5pm, Rm 508 in building 201E (Human Sciences)
- ▶ Five assignments (20%)
- ▶ Mid-term test (20%)
- ▶ Final exam (60%)

Topics

1. Data mining overview
2. Data mining tools for prediction and classification
3. Unsupervised learning
4. Visualisation
5. Data wrangling
6. Handling large data sets

Useful Resources

Use the following resources supplement the lectures:

- ▶ The Elements of Statistical Learning
- ▶ Introduction to Statistical Learning
- ▶ R documentation
- ▶ Google
- ▶ Other books and papers, see references at the end of lectures
- ▶ Most of the references are available electronically
- ▶ There is an enormous amount of good stuff available over the net - Google madly!

Web page

<https://www.stat.auckland.ac.nz/~lee/784/>

See also Thomas's page at

<https://www.stat.auckland.ac.nz/~yee/784/>

- ▶ Course information
- ▶ Notice board - check frequently!
- ▶ Lectures
- ▶ Past exams and assignments

Web page

The screenshot shows a web browser displaying the 'STATS 784 Homepage' for the Department of Statistics at the University of Auckland. The page features a navigation menu on the left with links for Home Page, Notices, Course Information, Lectures, Assignments, Lab Sheets, Getting R On Your Computer, Resources, and Other Useful Links. The main content area is titled 'Welcome to the STATS 784 Homepage' and includes a brief description of the page's purpose. It lists the lecturers as Alan Lee and Thomas Vee, accompanied by their respective headshots. Contact details for both lecturers are provided at the bottom of the main content area.

STATS 784 Homepage



UNIVERSITY OF AUCKLAND
DEPARTMENT OF STATISTICS

SCIENCE Home | About | Courses | People | Consulting | Research | Grants | Links

Welcome to the STATS 784 Homepage

This web page is designed to keep you informed about the course. It contains links to resources you may find useful in working through your topics. To supply feedback on any aspect of the course, including those web pages, email either [Alan](#) or [Thomas](#).

Taught by:



Alan Lee **Thomas Vee**

Contact details:
Alan - Stop by my office, Rm 205, Building 303S, call me on 373 7599 Extn 88768 or send me an email.
Thomas (after mid-semester break) - Stop by my office, Rm 312, Building 303, call me on 933 8811 or send me an email.

Class Rep

Any volunteers???

Today's agenda

Now we begin our discussion of data mining. Today we will cover

- ▶ What is data mining? Why is it increasingly important?
- ▶ The rise of Big Data
- ▶ The data mining process
- ▶ Predictive analytics, machine learning, supervised and unsupervised learning
- ▶ What we will study in the course

What is data mining: some definitions

“The extraction of implicit, previously unknown and potentially useful information from data”

Professor Ian Witten, author of “Data mining : practical machine learning tools and techniques”

“The process of discovering useful patterns in large data sets”

Daniel Larose, author of “Discovering Knowledge in Data”

“From a statistical point of view, data mining is the computer automated exploratory data analysis of (usually) large complex data sets”

Jerry Friedman, co-author of “The Elements of Statistical Learning”

More definitions

“Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.”

Wikipedia

“Data mining is the process of extracting previously unknown, valid, and actionable information from large data bases and using it to make crucial business decisions”

Alex Zekulin, former head of IBM's data mining group.

.... and so on

Common themes

- ▶ Discovery of previously unknown patterns
- ▶ Discovery of useful information
- ▶ Large data sets
- ▶ Computerised
- ▶ Leads to better decisions

Why is there so much hype?

- ▶ Rise of computer technology has enabled the collection and storage of vast amounts of information - point of sale transaction data (supermarkets), internet transactions, Google, Facebook, Twitter, cellphone data, tax data, and so on - Big Data
- ▶ Commercial organizations (and others) are trying to mine this data deluge for competitive advantage
- ▶ Hence the rise of Data Science as a discipline
- ▶ Many employment opportunities for those with with skills in statistics and computing - see the article "Data Scientist: The Sexiest Job of the 21st Century" by Thomas H. Davenport and D.J. Patil in the October 2012 issue of the Harvard Business Review

Industries of the future?

According to Alec Ross (former advisor to Hillary Clinton) the industries of the future include robotics, cybersecurity, big data analytics and genomics. See his book "The Industries of the Future", showcased in the Listener, 25th June issue



Big Data Timeline

1991 : Internet is born

1995 : GPS becomes fully operational

1998 : Google founded

2001 : Wikipedia founded

2003 : Data created in 2003 surpasses that all created in human history up to 2002

2004 : 1 million articles on Wikipedia, Facebook founded

Big Data Timeline (cont)

2007 : iPhone

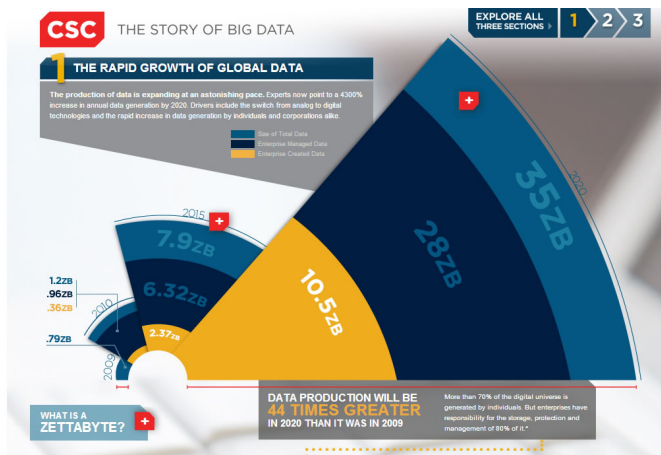
2008 : Number of devices connected to the internet exceeds worlds population

2011 : No more room in IPv4 address space (2^{32})

2012 : Less than 3% of data being analysed, each office worker generating 5gB of data per day

2014 : Adoption of cloud ERP (enterprise resource planning, business information system)

Growth of big data



Data mining projects

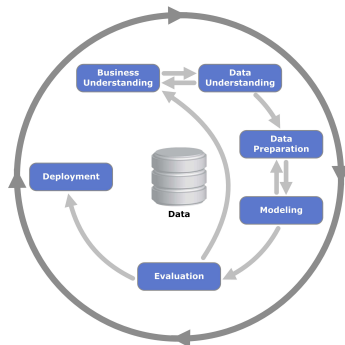
A typical data mining project has the following stages

1. A question is floated by senior management (Can we get a better forecast of next year's milk production? How can we better manage the replacement schedule for the city's infrastructure?)
2. Data requirements are scoped out, data sources identified
3. Data is assembled, cleaned and prepared for analysis
4. Model building, construction of predictions, identification of relevant patterns
5. Selling management
6. Implementation

The term “Data Science” refers to stages 2 to 4, “Data Mining” more to the task 4.

CRISP-DM

Another version of the previous slide: *Cross Industry Standard Process for Data Mining*. See the Wikipedia article in the references for more information.



See also the SAS version SEMMA (Sample, Explore, Modify, Model, and Assess).

Multidisciplinary skills

These stages require complementary skills

1. Business knowledge to identify projects, identify relevant data
2. Computing/data base skills to assemble the data into the correct form for analysis
3. Statistical skills to fit models, make predictions, identify patterns
4. Communication skills to translate/sell findings and recommendations to management

An example: Fonterra milk supply forecast

The problem: to make the forecasting of the season's milk supply more transparent. Existing forecast was based on subject matter knowledge, expressed in spreadsheets, very dependent on a small team.

1. Meeting of management and consultants, problem scoped out.
2. Subsequent meeting to identify data sources, acquire subject matter knowledge
3. Model fitting, validation against previous forecasts, documentation
4. Presentation and implementation.

The data science tasks

1. Data Wrangling (can be up to 80% of total effort)
 - ▶ Data acquisition
 - ▶ Interaction with data bases, internet, API's (application programming interface)
 - ▶ Synthesis of data from various sources
 - ▶ Imputation and cleaning
2. Data Mining (statistical learning, data analysis, modelling)
 - ▶ Exploratory analysis
 - ▶ Feature selection/engineering
 - ▶ Identification of patterns (association rules, unsupervised learning)
 - ▶ Making predictions, classifying (supervised learning)

Data Science refers to both of these tasks.

DM: Connection with traditional statistics

- ▶ Differences
 - ▶ Algorithms not mathematical theory, empirical performance measures
 - ▶ Little concern with probability models for errors
 - ▶ Many techniques (CART, neural networks, boosting) originated outside statistics, in computer science, artificial intelligence
- ▶ Similarities and convergence
 - ▶ Overlapping techniques (linear methods, clustering, prediction, classification)
 - ▶ Importance of computing and visualization
 - ▶ Statistical interpretations of DM techniques (Friedman et al.)
 - ▶ Inclusion in mainstream statistical education, statistics becoming more “DM-like”
- ▶ See Jerry Friedman’s article “Statistics and data mining: What’s the connection”

Statistical learning

The extraction of information from data relies heavily on a group of techniques:

- ▶ Visualization
- ▶ Prediction and classification
- ▶ Finding patterns using clustering and association rules

Collectively, these are tools for statistical learning. *Supervised learning* refers to prediction and classification, where we want to predict the value of a variable (target, response) in terms of other variables (features, predictors, covariates). It is based on models, where we assume some mathematical relationship between the target and the features.

Unsupervised learning is when we don't differentiate between target and features, but want to identify patterns and understand structure. Based on clustering, dimension reduction, visualization, association rules.

DM: some caveats

- ▶ When making predictions, we implicitly or explicitly assume that the future data being predicted is in some sense similar to the data used to construct our prediction rule. In traditional statistics, we assume that the distribution of the data in these two phases (construction and prediction) is the same. If this is not the case, our predictions will be biased, no matter how big our data sets are. We cannot rely on the “bigness” of big data to protect us from error.
- ▶ When finding patterns in data, we must beware of spurious patterns in the data arising by chance or by some artifact in the the data collection process - the phenomenon of typing monkeys producing the works of Shakespeare. This problem is made worse by the size of modern data sets.

Data

Data can assume many forms: data base tables, twitter feeds, images, emails, spreadsheets, SAS data sets. it can be in many formats (raw text, XML, JSON) Most statistical learning procedures require these data are processed into rectangular data sets acceptable to the DM software, with rows representing measurements on objects, and columns representing variables (the measurements). Often no “missing values” can be present.

Getting data into this form is a very big part of data science.

Computers are getting more powerful, but data is getting bigger. Our rectangular data set may not fit into computer memory, or take too long to process. Thus, parallel computing, segmentation of computations, data streaming (work on a bit at a time) are important data science topics.

Data Cleaning and Imputation

Much real-world data is “dirty” full of errors and missing values. Data cleaning (correcting of errors) and imputation (filling in of missing values) is really important.

Imputation will be covered in some detail in a later lecture.

Software

Many DM software packages available, some free!

- ▶ R: free, comprehensive, increasing ability to handle large data sets, some gui support (Rattle).
- ▶ SAS Enterprise Miner: Comprehensive, GUI interface, expensive.
- ▶ WEKA(Waikato Environment for Knowledge Analysis): A suite of software from Ian Witten's group at Waikato, has an R interface. See Ch 10 of Ian's book "Data mining : practical machine learning tools and techniques".
- ▶ ... and 000's of others.

What's next

We have seen that the data mining process involves many different skills. This course concentrates mainly on the “mining” (statistical learning) aspect, emphasizing the techniques used to extract useful information from data. This will be the subject matter of the first half of the course, which will be mainly devoted to predictive analytics (supervised learning).

In the second half, Thomas will continue the “mining” theme, covering unsupervised learning. He will also cover visualization techniques, and spend some time on the data acquisition/data wrangling/data cleaning aspects. Data wrangling is also covered in more detail in the data science course STATS 769.

References

1. Davenport, T and Patil, D. (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, 90, pp 70-76.
2. Dean, J. (2014). Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. Wiley.
3. Friedman, J. (2007). Data mining and Statistics: Whats the connection? <http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf>
4. Hastie, T., Tibshirani, R.J. and Friedman, J. (2009). The Elements of Statistical Learning, 2nd Ed. Springer.
5. G James, D Witten, TJ Hastie and RJ Tibshirani. (2013). An Introduction to Statistical Learning. Springer.
6. Kuhn, M. and Johnson, K. (2013). Applied Predictive Modelling. Springer
7. Larose, D and Larose C. (2014). Discovering knowledge in data : an introduction to data mining, 2nd Ed. Wiley.
8. Ross, A (2016). The Industries of the Future. Simon and Schuster.
9. Wikipedia (2016) Cross Industry Standard Process for Data Mining, https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
10. Witten, I., Frank E., and Hall, M. (2011). Data mining : practical machine learning tools and techniques, 3rd Ed. Morgan Kaufmann.