STATS 769
Overview

Paul Murrell
Alan Lee

The University of Auckland

July 21, 2015

Overview

- This course covers topics related to the analysis of large and/or complex data sets.
- We will cover a range of modern data mining techniques.
- We will cover a range of data technologies for accessing and processing data.

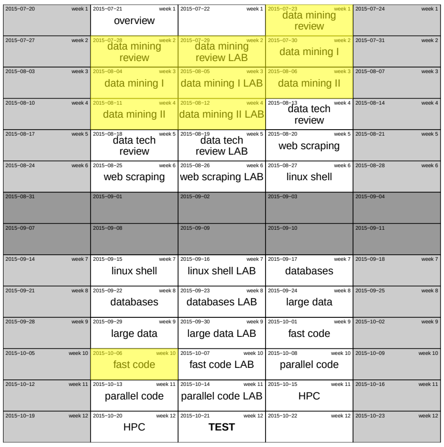
> *In my view, we do need a term that covers that full life cycle - not building an IT platform and network for databases perhaps, but yes, building the occasional database given the infrastructure is in place; managing large amounts of data and knowing enough about IT at least to know where and how to get help; reshaping and wrangling many many datasets and knowing the tricks of the trade to combine them in meaningful ways; knowing how to manage complex, large, conflicting systems of classifications; fitting statistical models and performing inference; and efficiently presenting results in high quality presentations and graphics using tools like LaTeX, JavaScript, Shiny, etc.*

Peter Ellis
MANAGER SECTOR PERFORMANCE
Ministry of Business, Innovation & Employment

Course structure

- Lecture Tuesday 9-10 206.201
- Lab Wednesday 8-10 303.175
- Lecture Thursday 8-9 303.B07

- There is no lab in the first week.
- We will spend one "week" on each topic: Thursday Lecture, Tuesday Lecture, Wednesday Lab.
- You will have a whole week to complete the Lab work (not just the 2-hour Lab session) and the Lab work will be assessed.
- There is a TEST, instead of a lab, in the last week.



2015−07−20 week 1	2015−07−21 week 1 overview	2015−07−22 week 1	2015−07−23 week 1 data mining review	2015−07−24 week 1
2015−07−27 week 2	2015−07−28 week 2 data mining review	2015−07−29 week 2 data mining review LAB	2015−07−30 week 2 data mining I	2015−07−31 week 2
2015−08−03 week 3	2015−08−04 week 3 data mining I	2015−08−05 week 3 data mining I LAB	2015−08−06 week 3 data mining II	2015−08−07 week 3
2015−08−10 week 4	2015−08−11 week 4 data mining II	2015−08−12 week 4 data mining II LAB	2015−08−13 week 4 data tech review	2015−08−14 week 4
2015−08−17 week 5	2015−08−18 week 5 data tech review	2015−08−19 week 5 data tech review LAB	2015−08−20 week 5 web scraping	2015−08−21 week 5
2015−08−24 week 6	2015−08−25 week 6 web scraping	2015−08−26 week 6 web scraping LAB	2015−08−27 week 6 linux shell	2015−08−28 week 6
2015−08−31	2015−09−01	2015−09−02	2015−09−03	2015−09−04
2015−09−07	2015−09−08	2015−09−09	2015−09−10	2015−09−11
2015−09−14 week 7	2015−09−15 week 7 linux shell	2015−09−16 week 7 linux shell LAB	2015−09−17 week 7 databases	2015−09−18 week 7
2015−09−21 week 8	2015−09−22 week 8 databases	2015−09−23 week 8 databases LAB	2015−09−24 week 8 large data	2015−09−25 week 8
2015−09−28 week 9	2015−09−29 week 9 large data	2015−09−30 week 9 large data LAB	2015−10−01 week 9 fast code	2015−10−02 week 9
2015−10−05 week 10	2015−10−06 week 10 fast code	2015−10−07 week 10 fast code LAB	2015−10−08 week 10 parallel code	2015−10−09 week 10
2015−10−12 week 11	2015−10−13 week 11 parallel code	2015−10−14 week 11 parallel code LAB	2015−10−15 week 11 HPC	2015−10−16 week 11
2015−10−19 week 12	2015−10−20 week 12 HPC	2015−10−21 week 12 **TEST**	2015−10−22 week 12	2015−10−23 week 12

Assessment

- Each Lab will be worth 3% of your mark (for a total of 30%).
- There will be one assignment worth 30%.
- There will be a written test (in the final Lab session) worth 40%.
  You must pass the test in order to pass the course.
- There is no exam.

Resources

- "An Introduction to Statistical Learning"
  Gareth James,Daniela Witten, Trevor Hastie and Robert Tibshirani
  `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf`
- "Introduction to Data Technologies" (assumed knowledge)
  Paul Murrell
  `https://www.stat.auckland.ac.nz/~paul/ItDT/`
- "Advanced R"
  Hadley Wickham
  `http://adv-r.had.co.nz/`
- "XML and Web Technologies for Data Sciences with R"
  Deborah Nolan and Duncan Temple Lang
  `http://link.springer.com/book/10.1007%2F978-1-4614-7900-0`

Class Rep

- Volunteer(s) required !

Data mining topics

Two main topics

- Supervised learning (aka prediction, classification) 4 lectures
- Unsupervised learning (aka clustering, ordination) 2 lectures

Examples

- The California housing data: data on 20640 census areas ("block groups") Task is to predict average house price from 9 geographic and economic covariates measured at the census area level. See `https://www.stat.auckland.ac.nz/~lee/760/houses.txt`
- The spam data: 59 variables measured on 4096 email messages, classified as spam or genuine. The aim is to develop a classification rule to decide if a new email is spam or genuine. Used in the design of spam filters. The variables relate to word frequencies and aspects of punctuation and capitalization.See `http://statweb.stanford.edu/~tibs/ElemStatLearn/index.html`

Supervised learning

Here we use data to make predictions about the future.

> *It's tough to make predictions, especially about the future.*

Yogi Berra, New York Yankees

Prediction is done by fitting "regression" models

$$y = f(x_1, \ldots, x_k) + \text{error}$$

y : Quantity to be predicted, (response, output, target)

f : Some function, e.g.linear, spline, tree etc

x's : Other variables (explanatory variables, inputs, features) used to predict

Supervised learning: tasks

- Choose function (predictor) $\hat{f}$ from some set of functions, hopefully a rich flexible set. This will be an approximation to the true f, hopefully a good one.
- Choose inputs, from an available set. Hopefully this includes the most important ones.

Supervised learning: training data

Most modelling software expects data in a rectangular "cases by variables" form (training data)

	x_1	x_2	$\cdots$	x_k
case 1	.	.	$\cdots$	.
case 2	.	.	$\cdots$	.
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
case n	.	.	$\cdots$	.

Supervised learning: training data (cont)

- Training data is used to select the predictor and the required inputs (fitting the model, training the predictor). Not all inputs in the training data will necessarily be used in the predictor.
- Creating the data in this form is a major task and is the subject matter of most of the course. There may be substantial work required to "clean" the data (correct errors, impute missing values), often 80% or more of a whole project.

Supervised learning: predicting

Having fitted the model, given new data $x_1, \ldots, x_k$ we predict the corresponding value of y by

$$\hat{f}(x_1, \ldots, x_k).$$

Most R data mining software will have two parts:

- A "fitting function" to fit the model e.g. `lm`, and
- A function (usually the generic function `predict`) to calculate the predictions from new data.

Supervised learning: Classification

If the response y is a category rather than a numerical quantity, we sometimes speak of "classification" rather than "prediction", and call $\hat{f}$ a "clasification rule".

Most data mining methods can be tweeked to perform both prediction of numerical quantities and classification with only minor changes in the method, using essentially the same software for either type of task.

Unsupervised Learning

- Given a set of individuals/objects, can we group them into “clusters” of similar objects? (Called “segmentation” in market research, where we divide customers into “segments” and pitch separate advertising to each segment).
- More generally, can we visualise the structure of the data and the relationships between objects? (ordination, reduction in dimension, so we can draw 2 or 3-dimensional pictures of high dimensional data)

Data technology topics

- Coping with data that is not already rectangular.
- Coping with data that is too large
  (for RAM or your hard drive).
- Making your code run faster.

Non-rectangular data

- Data from the Web.
- NoSQL databases.
- Data formats (JSON and XML).

Large data

- Make the machine bigger.
- Make the data smaller.
- Use different software.

Faster code

- Writing fast (and slow) R code.
- Making use of multiple cores.
- Making use of remote machines.