FIT1043 Introduction to Data Science

Module 4

Data Resources, Processes, Standards and Tools

Lecture 7

Monash University

Discussion: Unix Shell

Useful for managing and manipulating **large files**

- **without ever loading them fully into memory**
- using pipes allow us to process files as a stream
- allows us to deal with files that are too big for applications and/or don't fit into memory

Shell contains many useful commands, like

- less to view large files
- grep to search large files
- awk to process them one line at a time (and cut them down to size for visualising)

Discussion: Hadoop

Hadoop provides an inexpensive and open source platform for parallelising processing

- based on simple Map-Reduce architecture
- broad range of tools and easy to use
- not suited to streaming or where a pipeline architecture is needed
- perhaps approaching Plateau of Productivity phase (in hype cycle)

Discussion: Spark

Google's dataflow and Spark's DAG processor are more recent developments (than Hadoop)

- include Map-Reduce capabilities
- but also provide for shared-memory
- useful for training Machine Learning models in a distributed fashion
- Google's dataflow can be mapped to Spark's DAG
- Further information:
  - *"Will Spark, Google Dataflow Steal Hadoop's Thunder?"* on *Information Week*
  - *Pivotal blog on real-time data and Spark growth*

Unit Schedule: Modules

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	**7**	**resources and standards**
	8	**resources case studies**
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	data management frameworks

Introduction to Resources (ePub section 4.1)

introduction to issues

- using data
  - new data sources or clever and creative use / combination of multiple existing data sources
- open data
  - organisations provide machine readable data to support data science
- wrangling
  - manipulating data to make it directly usable for analysis
- standards
  - support cooperation, reuse, common tools, *etc.*

Introduction to Resources
Using data

access to new data sources or clever and creative use of existing multiple data sources are hallmarks of innovative data science

Four Examples of Using Data

We'll now look at four examples of public data and using data?

1. NYC data
2. traffic prediction
3. web pharmacovigilance
4. predictive analytics for banks

New York City Data

Under Mayor Bloomberg, NYC embarked on a program to make the city's data accessible:

- *"How data and open government are transforming NYC"* in *Radar.O'Reilly*:
  - "In God We Trust," tweeted New York City Mayor Mike Bloomberg this month. "Everyone else, bring data."
  - applications of the data provided:
    - "real-time updates on your phone based on where the buses are located using very low-cost technologies"
    - applying predictive analytics to building code violations and housing data to try to understand where potential fire risks might exist
- *Bloomberg signs NYC 'Open Data Policy'* into law, plans web portal for 2018," in *Engadget*
- *NYC Open Data portal*
- Melbourne has a similar portal: *City of Melbourne's open data platform*

NYC Data, cont.

"How we found the worst place to park in New York City" is examples, and a discussion of the complexities of getting data out of NYC:

Map of road speed by day+time: GPS data for NYC cabs gives; data obtained via FOIL (Freedom of Information Law) request, then made public by recipient

Danger spots for cycles: *NYPD crash data* obtained by daily download of PDF files followed by (non-trivial) extraction

Dirty waterways: *fecal coliform measurements on waterways* from Department of Environmental Protection's website; extracted from Excel sheets per site; each in a different format

Faulty road markings: parking tickets for fire-hydrants by location from *NYC Open Data portal* need to normalize the addresses supplied

Traffic Prediction

see 7:40-11:06 on Clearflow in
"Data, Predictions, and Decisions in Support of People and Society,"
by Eric Horvitz

- forecasting traffic: blockages, clearing, surprising situations, alternate routes
- critical data:
  - GPS data on traffic flow
  - maps
  - incidents and events
  - weather
- see *Microsoft Introduces Tool for Avoiding Traffic Jams* in *NYT* 2008

Drug Interactions

pharmacovigilance ::= monitoring the effects of medical drugs after they have been licensed for use

Example:

- see 38:40-42:30 on in *"Data, Predictions, and Decisions in Support of People and Socie*
  by Eric Horvitz
- **prediction task** is to tell which drug pairs interact to cause hyperglycemia

Drug Interactions, cont.

Using user Web queries to determine which drugs cause interactions.

- *FAERS* gathers data from physicians about drug interactions observed in their patients
- training data comes from the Web (FAERS data confirming which pairs cause interaction)
- test data is drug pairs and the reporting ratio (RR) computed from web queries (example data below is fictional)

drug 1	**drug 2**	**RR**	**truth**
dobutamine	hydrocortisone	12.6	causes
glipizide	phenotoin	9.4	causes
...	...	...	...
budesonide	formoterol	7.3	not
labetalol	sertraline	2.4	not

Drug Interactions, cont.

drug 1	drug 2	RR	truth
dobutamine	hydrocortisone	12.6	causes
glipizide	phenotoin	9.4	causes
...	...	...	...
budesonide	formoterol	7.3	not
labetalol	sertraline	2.4	not

- choose a `cut-off` for **RR** and when (RR>`cut-off`) we predict “causes”, otherwise “not”
- changing the `cut-off` lets us control the accuracy of our predictions
- either, more accurate but small coverage
- or, less accurate and larger coverage

Prediction Outcomes

- so (RR>cut-off) corresponds to "prediction=causes"
- change the `cut-off` for **RR** and see how it affects our predictions
- for a given `cut-off`, we can fill out a table, where the counts are the entries

	truth=not	truth=causes
prediction=not	19	7
prediction=causes	12	24

- true positive ::= entry for "prediction=causes" and "truth=causes", false negative ::= entry for "prediction=not" and "truth=causes", *etc.*
- true positive rate ::= $\frac{\text{true pos.}}{\text{true pos.+false neg.}} = \frac{24}{24+7}$
- false positive rate ::= $\frac{\text{false pos.}}{\text{false pos.+true neg.}} = \frac{12}{12+19}$
- see *Sensitivity and specificity*

ROC Curves

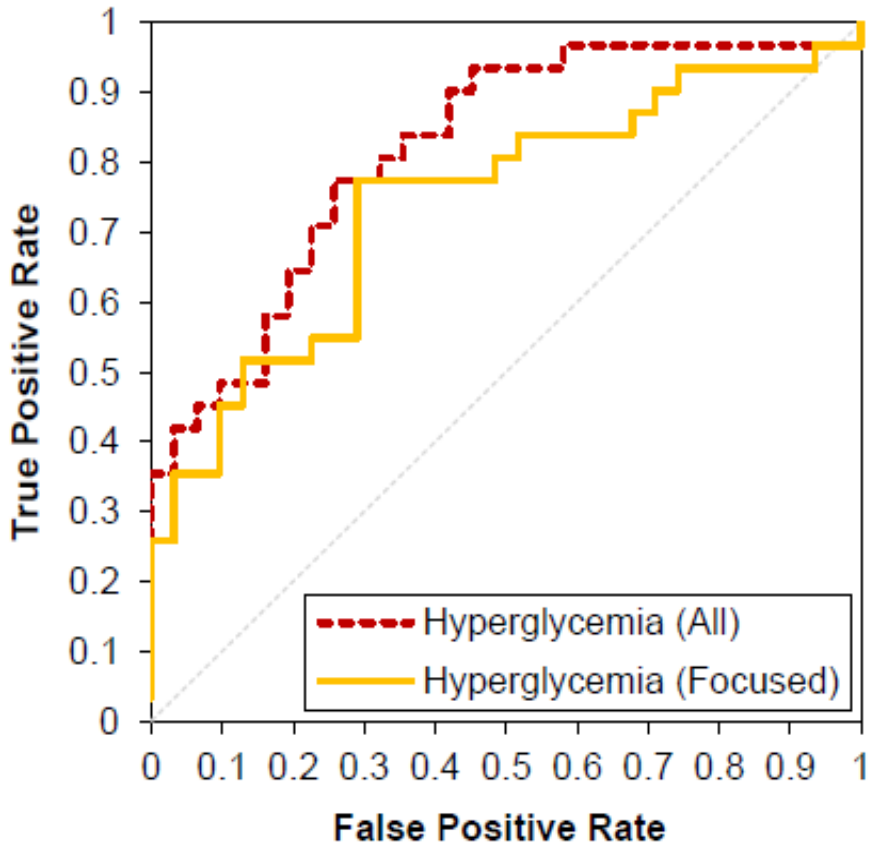
- now for each cut-off we have a point in 2-D (true positive rate,false positive rate)
- as we increase the cut-off true positive rate monotonically increases in steps; and false positive rate monotonically decreases
- plotting the values yields the so-called ROC curve (ROC is *receiver operating characteristic*)
- ROC curve answers question:

  "how do the error rates change as I change my cut-off?"

Characterizing Sensor Error

Test on known interactions

- 31 true positives for hyperglycemia
- 31 true negatives for hyperglycemia



Label	*Drug 1*	*Drug 2*
TP	dobutamine	hydrocortisone
TP	dobutamine	triamcinolone
TP	dobutamine	prednisolone
TP	betamethasone	dobutamine
TP	glipizide	phenytoin
TP	dobutamine	methylprednisolone
TP	prednisolone	salmeterol
TP	salmeterol	triamcinolone
TP	betamethasone	terbutaline
TP	dexamethasone	dobutamine
TP	budesonide	salmeterol
TN	hydrochlorothiazide	tazobactam
TN	clindamycin	montelukast
TN	lamotrigine	nystatin
TN	methylprednisolone	rosuvastatin
TP	budesonide	formoterol
TN	loratadine	nystatin
TN	hydroxychloroquine	prochlorperazine
TN	labetalol	sertraline
TN	ciprofloxacin	vecuronium

Web Pharmacovigilance

- so previous curves show web query data can be used quite reliably to predict drug interactions causing hyperglycemia, *i.e.* with need for obtaining the physician data
- the web query data is acting as *proxy data* for the FAERS reports
- you could almost perform the RR computations yourself using Googles results estimates!

Predictive Analytics for Banks

Foster Provost

- coauthor of the O'Reilly best-selling book, *Data Science for Business*.
- presented at Stata+Hadoop in 2013, *"Predictive Analytics with Fine-grained Behavior Data"*.
- The *video is here*.

Is big data better?

- Answer: Not always.
- But big data can be better when its richer, more fine-grained.

Introduction to Resources
Open data

organisations provide machine readable to support data science

Start with the video *The year open data went worldwide* a TED talk by Prof. Sir Tim Berners-Lee (video, 6 mins)

Democratization of Data

"The New Data Republic: Not Quite a Democracy" in MIT Sloan Review 2015

- from Hal Varian: "information that once was available to only a select few. . . available to everyone"
- from Robert Duffner: "finally puts crucial business information in the hands of those who need it"
- government and IT departments building data and infrastructure to allow sharing
  - *USA Open Gov Initiative*
- analytic tools, desktop and web-based, available to analyse it
- but people need the right skills

open data is all good and well, but people need to be able to use it too!

Open Data Recommendations

"Open data: Unlocking innovation and performance with liquid informa
by MGI and *"Science as an open enterprise"* report by the Royal Society (UK) claim:

- open data provides new opportunities for business, new products and services, and can raise productivity
- open data supports public understanding and citizen engagement
- scientists need to better publicise their data (with help from universities, *etc.*)
- industry sectors should work with regulators and coordinate industry collaboration
- collaboration across sectors in both public and private settings, *e.g.*, disaster response, education

Linked Open Data

LOD project started by
Prof. Sir Tim Berners-Lee, OM, KBE, FRS, FREng, FRSA, DFBCS.

- objects given a URI (like a URL)
- relationships between two objects can be represented as a triple, (subject, verb, object)
- relation itself is another URI
- data has an open license for use

e.g. *NYT* or *Eighth Avenue in Manhattan*

- a *tutorial on LOD* by Tom Heath

Introduction to Resources
Wrangling

manipulating data to make it directly usable for analysis

Wrangling Examples

Examples of wrangling tasks:

- extract the core news text, title, and date from a webpage:
  Apple's iPhone loses top spot to Android in Australia
- extract the text plus details from a PDF file:
  "Data Wrangling: The Challenging Journey ..."
- extract all article titles from an XML file:
  PUBMED results xml
- digitize the text from a scanned image:
  scanned letter
- extract all the sentences referring to particular individual in an article:
  a news article about Hillary Clinton

Wrangling Examples (cont.)

More wrangling tasks:

- integrate data sources:
  company has customer records in 4 different databases in different formats; you want a single standardised set of customer names and addresses
- geocoding:
  convert addresses in your customer database into geographic latitude and longitude
- convert free text dates to standard format:
  e.g. map: “next Tuesday” $\rightarrow$ “2nd January 15”,
  other date examples: “January 3 next year” , “3rd Friday in the month” “03/31/15”, “31/03/15”

Wrangling Examples (cont.)

More wrangling tasks:

- recognise missing values and deal with them, by e.g.
  - removing the row or column,
  - replace with a special “unknown” value,
  - replace with an average value,
  - or doing nothing
- deal with outliers or “illegal” values,
  e.g. remove extremely large values that are likely due to sensor noise
- discretise the data into a set of values
  discretisation is necessary if the predictive model being learnt cannot handle continuous data

Introduction to Resources
Standards

support cooperation, reuse, common tools, *etc.*

Example Standards

Examples of standards

- Metadata standards
  - such as *Dublin Core*
- XML formats for sharing models,
  - e.g. PMML (see below)
- Standards for describing the data mining/science process,
  - such as *CRISP-DM*
- Standard vocabularies for use in Medicine, e.g.
  - health codes: disease and health problem codings *ICD-10*
  - systematized nomenclature of medicine, clinical terms, *SNoMed-CT*

What other sorts of things might you have standards for?

Standards and Issues (ePub section 4.5)

more on standards and issues

- some standards
  - some standards for semi-structured data, data science process and predictive models
- open data and open source software
  - critical infrastructure and tools
- APIs and SaaS
  - think Web 3.0

Standards and Issues

Some standards

some standards for semi-structured data, data science process and predictive models

Data Science Process

We've seen many data science processes and lifecycles:

- e.g. our own "standard Data Science value chain"
- *CRISP-DM* discussed previously, is a standardised data science process
- statisticians sometimes use the term exploratory data analysis for part of the process

Semi-Structured Data

Semi-structured data is data that is presented in XML or JSON:

- see some examples for *here*
- Note YAML (Yet Another Markup Language), which is just an indentation (easier to read) version of JSON
- standard libraries for reading/writing/manipulating semi-structured data exist in Python, Perl, Java
- don't need to know all the details of XML (and related Schema languages)
  many good online tutorials, *e.g.* *W3schools.com*
- their use in systems leads to the open world assumption about data, where we may download relevant data on the fly from APIs *etc.*

Model Language

PMML ::= Predictive Model Markup Language

PMML provides a standard language for describing a (predictive) model that can be passed between analytic software (e.g. from R to SAS).

- *PMML: An Open Standard for Sharing Models*
- A list of products working with PMML is the *PMML Powered page* on DMG site.

Unit Schedule: Next Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	**resources case studies**
5.	9	data analysis theory
	10	data analysis process
6.	11	issues in data management
	12	data management frameworks