FIT5145 Introduction to Data Science

End of Semester

Summary

2019 Lecture 12

Monash University

Reminders

- SETU time: see SETU Unit Evaluation link in Moodle
- Reminders:
  - You should have handed in your Data Case-study Report and Presentation already
  - This week you'll present the presentation during your tutorial

Unit Schedule: This 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	**GUEST SPEAKER & EXAM INFO**

Home Activity: Privacy and Security

Investigate issues related to security and privacy of data using (On Moodle under Tutorial resources week 12):

- Legal requirements for companies dealing with sensitive user data.
- Example of private data (ENRON email corpus)
  - Very easy (with a couple of shell commands) to discover very sensitive information (mobile phone numbers, credit card information, etc.)
- Famous information leaks
  - Some very scary leaks ....
- Example website privacy policies:
  - What information is Google storing about you?
  - Why are they keeping that information?
  - What control do they provide you with over the information they collect.

Guest Speaker with Q&A

- Mr. Salim Naim
  - CTO Advance Analytics & Data Science Microsoft Services, APJ
  - On 25th October, 10am-12pm
  - Location, Lecture theatre K309 Building K

Microsoft

The Exam

- Content of the Exam
    - What is examinable?
- Format of the Exam
    - What will the exam paper look like?

Content of the Exam

What is examinable?

- Everything discussed in the lectures is examinable.
- That includes the "Brief Introduction to ..." slides:
  - on Python, R, Unix Shell, Decision Trees
  - **but** you do not need to memorise all the syntax for the programming languages!
- Content linked from lecture slides is not **directly** examinable
  - i.e. you **do not** need to learn everything that is linked from the lecture slides (there is a huge amount of content)
  - **but** sometimes the definitions/explanations of the content discussed in the lectures *is given in the linked content*,
  - so you will have had to follow the links (watched the video or skimmed the blog posts, etc.) to understand the lecture material properly!

Content of the Exam (cont.)

What is examinable?

- Content on Alexandria provides a very useful description of the content of the course
  - so most of it is examinable
  - reading it also provides a very useful revision tool!
- Content of the tutorials explains concepts from the slides
  - so it is examinable
  - but you don’t need to rote learn syntax!

Format of the Exam

What will the exam paper look like?

- Exam consists of two parts:
  - 42 multiple-choice questions (worth 42% of total mark)
  - 29 short-answer questions (worth 58% of total mark)
- Duration 2 hours
- Closed book
- No need to bring a calculator
- Sample questions available in SAQs, in lecture slides, and on Moodle (later) etc.

Unit

So, what did we cover in this unit?

- Quick overview of what we learnt

Week 1

- What is data science?
- What is machine learning?
- What is big data?
- Data science process and data science value chain
- Introduction to Python for data science

Week 2

- What does a data scientist do?
- What skills do they need?
- Impact data science is having
    - cloud services, effect on science, social good
- Tutorial
    - Investigated Motion charts as a data visualisation tool
    - Getting familiar with Python
- Home activity
    - Jobs in data science

Week 3

- Data business models
- Analytics levels: Descriptive, Predictive and Prescriptive Analytics
- Modeling decision problems with Influence Diagrams
- Data business models:
  - information brokering services
  - information-based differentiation services
  - information-based delivery network services
  - data providers
- Introduction to Python for data science (part 2)
- Tutorial
  - Getting more familiar with Python

Week 4

- Data science case studies
- Characterising them in terms of:
  - data sources
  - data volume, velocity, variety, veracity
  - software, analytics, processing
  - security, privacy
- Introduction to R for data science
- Tutorial
  - Modeling with influence diagrams

Week 5

- Characterising big data:
  - Volume, Velocity, Variety, Veracity, Variability, Visualisation, Value
- What is metadata?
  - different types of metadata
- Growth laws related to big data:
  - Moore’s law, Koomey’s law, Bell’s Law and Zimmerman’s Law
- Introduction to Unix Shell commands for data science
- Tutorial:
  - Exploratory analysis of big data in R

Week 6

- Processing big data
  - different types of databases (SQL, semi-structured, graph, noSQL, etc.)
  - different types of processing (interactive, streaming, batch)
  - distributed processing (map-reduce, spark, etc.)
- Tutorial:
  - Manipulating large files in the shell

Week 7

- Resources and the use of big data
- What is open data?
- What is data wrangling?
- Standards for publishing data and models
- Tutorial:
  - Understanding map-reduce

Week 8

- Common tools used (Hadoop and related Apache tools)
- APIs and Software-as-a-Service
- Case studies
- Tutorial:
  - Wrangling with SAS, DataWrangler and Python

Week 9

- Types of data analysis:
  - prediction, prediction with unknown variables, clustering, forecasting, etc.
- Learning theory
  - error vs loss functions
  - linear and polynomial regression
  - overfitting due to overly complicated model / insufficient data
  - training and test split
  - signal to noise
  - ensembling multiple models
- Tutorial:
  - Wrangling big text data (from Twitter) using shell commands

Week 10

- Correlation vs Causation and the need for controlled experiments
- Imputing missing values
- Examples of analytic software
- Case studies
- Introduction to Decision/Regression trees
- Tutorial:
  - understanding learning theory though examples in Python

Week 11

- Ethics and privacy
- Regulatory compliance
- What is Data Governance
- Data Management case studies
- Tutorial:
  - building predictive models with BigML

Week 12

- Home activity
  - Understanding Privacy, Legal Requirements and the Prevention of Information Leaks
- Guest lecture on Friday 25th October
- Phew! We’ve covered a lot of stuff in this unit!

THE END

- I hope you've learnt a lot from the unit
- Best of luck for your revision and the exam!