FIT1043 Introduction to Data Science

Unit Review and Exam Information

Reminders

- SETU time: see SETU Unit Evaluation link in Moodle
- Reminders:
  - Final assignment due 22 October 11:59pm
  - Use tutorial 6 and 8 shell commands to answer the questions
  - Copy/Paste your commands in your document too

Discussion: Privacy and Security

In week 12 tutorial we investigate issues related to security and privacy of data.

- 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?

The Exam

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

Content of the Exam

- Everything discussed in the lectures is examinable.
- That includes the "Brief Introduction to ..." slides:
  - on Python, R, Unix Shell
  - **but** you do not need to memorise all the syntax!
- Content linked from lecture slides is not **directly** examinable

  i.e. you **do not** need to learn everything that is linked too
  - **but** sometimes the definitions/explanations of the content discussed in the lectures *is given in the linked content*,
- Content on Alexandria provides a useful description of the content of the course
  - is not directly examinable, except where in slides
- Content of the tutorials explains concepts from the slides

Format of the Exam

What will the exam paper look like?

- Exam consists of two parts:
  - 15 multiple-choice questions (1 mark each)
  - 25 short-answer questions (2 marks each)
- For short-answer questions, your answer should be written in clear, simple English. Your answers should be complete enough in addressing the question. Extensive prose is not required. Structured bullet points are acceptable.
- Exam duration: 2 hours writing time
- Reading time: 10 minutes
- Close book
- No need to bring a calculator
- Sample questions available on Moodle ...

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

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
- Introduction to Python for data science
- Tutorial
  - Investigated Motion charts as a data visualisation tool
  - Getting familiar with Python
  - Jobs in data science

Week 3

- Data business models
- Analytics levels: Descriptive, Predictive and Prescriptive Analytics
- Modeling decision problems with Influence Diagrams
- Tutorial
  - Getting familiar with Python

Week 4

- Characterising big data:
  - Volume, Velocity, Variety, Veracity
- 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 R for data science
- Tutorial:
  - Modeling with influence diagrams

Week 5

- Processing big data
  - different types of databases (SQL, graph, noSQL, etc.)
  - different types of processing (interactive, streaming, batch)
  - distributed processing (map-reduce, hadoop, spark, etc.)
- Introduction to Unix Shell commands for data science
- Tutorial:
  - Exploratory analysis of big data in R

Week 6

- Resources and the use of big data
- What is open data?
- What is data wrangling?
- Introduction to predictive models
- Tutorial:
  - Manipulating large files in the shell
  - Understanding map-reduce

Week 7

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

Week 8

- 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
- Tutorial:
  - Wrangling big text data (from Twitter) using shell commands

Week 9

- Learning theory
  - training and test split
  - bias variance trade-off
  - ensembling multiple models
- Introduction to Decision/Regression trees
- Tutorial:
  - understanding learning theory through examples in Python

Week 10

- Imputing missing values
- Examples of analytic softwares
- Characterizing learning (Prediction?, Optimization? ..)
- Case studies
- Tutorial:
  - understanding learning theory through examples in Python

Week 11

- Confidentiality and privacy
- Regulatory compliance
- Data management
- Tutorial:
  - building predictive models with BigML

Week 12

- Guest lecture from Microsoft: Data and Artificial Intelligence- An Industry Perspective
- Tutorial:
  - Understanding Privacy, Legal Requirements and the Prevention of Information Leaks
- Phew! We've covered a lot of stuff in this unit!

THE END

- I hope you've enjoyed the unit
- Do consider follow-on units, where you'll learn the full stuff:
  - FIT2079 Data visualisation
  - FIT2086 Modelling for data analysis
  - FIT3152 Data analytics
  - more 3rd year units ...
- Best of luck for your revision and the exam!

"Our data analysis experts can't read your minds. You're going to reach your own decision regarding hiring us, even though it's the same decision we knew in advance you'd make."

"Sweetheart, my neural net predicts that you and I are 98.9% compatible. Will you be my Valentine?"