COMP20008 Elements of Data Processing:

Semester 2 2018

Lectures 24: Revision

Announcements

- Exam study guide has been updated to cover remaining material
- Sample exam and sketch answers is available

Plan today

- Project feedback
  - Phase 3
- Summary of subject
- Data science, what next?
- Exam stuff
- Reflections

Phase 3 Orals General Comments (1)

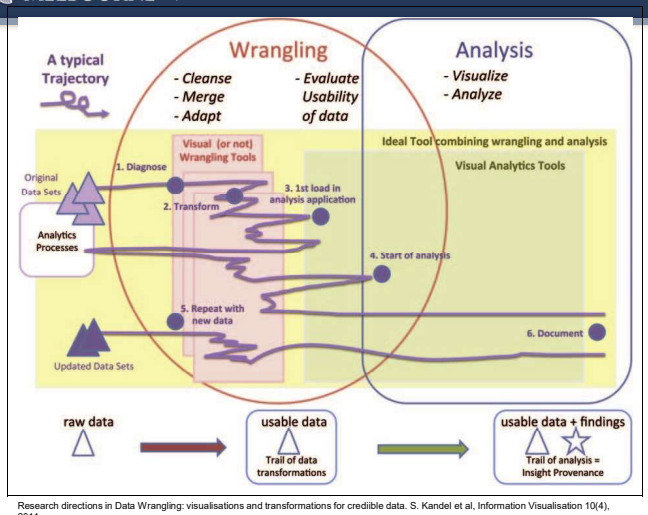
- Good points – oral
  - Generally interesting content and visuals
  - Fluent
  - Logical progression of ideas
  - Interesting designs (though minimal is least distracting)
  - Good degree of enthusiasm
- Good points - technical
  - Good use of different data analysis techniques to answer the chosen question in a logical manner.
  - Good dealing with the missing values and good explanation for the imputation methods

Phase 3 Orals General Comments (2)

- Points to monitor
  - Presentation
    - Avoid reading every word on slides
    - Pace and timing
    - Questions sometimes challenging
    - Eye contact
  - Slides
    - Outline was missing (structure of the presentation).
    - Motivation was key
    - Discuss/interpret figures
    - Direct audience where to look
    - Include readable figure labels/axes
    - Use 3 simple figures rather than 1 very complex figure

What has the subject been about?

- **Preprocessing** (4 lectures): Weeks 1-3
  - Data types and processing, data cleaning including outliers, missing data
- **Visualisation** (3 lectures): Weeks 3-4
  - Plotting and visualisation methods, clustering, dimensionality reduction
- **Analysis** (4 lectures): Weeks 5-6
  - Correlations, basic prediction techniques (decision tree, K-nearest neighbor)
- **Infrastructure and Distributed** (5 lectures): Weeks 7-9
  - Data linkage and data integration, blockchain
- **Social, privacy and ethics issues** (3 lectures): Weeks 10-12
  - K-anonymity, l-diversity, differential privacy, ethics

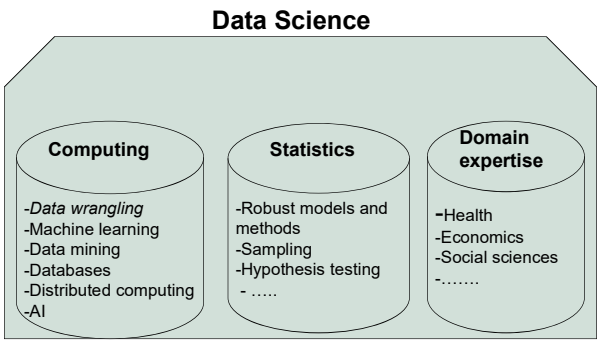


Research directions in Data Wrangling: visualisations and transformations for credible data. S. Kandel et al, Information Visualisation 10(4), 2011.

Projects: High level skills

- Using Python and pandas library
- Finding the right data, formulating an interesting question
- Wrangling data, getting into a usable format
  - Cleaning, imputation, outlier detection, integration
- Exploratory techniques for visualisation
- Basic methods for making predictions or assessing correlations
- Communication of findings

Want more?



Want more?

- Database Systems (INFO20003)
  - Relational databases
- Machine learning (COMP30027)
  - More sophisticated prediction/regression approaches
- Artificial Intelligence (COMP30024)
  - Robots/agents learning and operating in real environments
- Majors/degrees
  - Bachelor of Science (Computational Biology major)
  - Bachelor of Science (Data Science)
  - Master of Data Science

Want more? – Adding Breadth Cont.

- Complementary subjects
  - Subjects in statistics
  - Distributed/cloud/cluster computing
    - Apache/Hadoop/Hive/Pig, NoSQL
  - Develop specialist knowledge in a domain that interests you
    - Business, physics, chemistry, music, architecture&design, sport, education, biology …
    - Consider linkages to data science
  - Online courses through Coursera/Udacity/edX
- Other good stuff
  - Do an internship over the summer (e.g. Data61)
  - Participate in Kaggle (data prediction) competitions
  - Participate in Hackathons

Data Wrangling and data science careers

- Data scientist
  - “Sexiest job of the 21st century” (Harvard Business Review) ?
  - “No 1 job in the United States” (Glassdoor)
  - Financial Review May 23rd 2018
    - "A lot of people are transitioning from other fields like economics, psychology, mathematics, because they see the field is exploding and there's money to be made.”
    - "Experts in data science describe a wild west atmosphere right now, with little agreement over how to define the field and scores of people rushing to add "scientist" to their resumes whether or not it's accurate."

Good aspects

- Rewarding to see how data science results affect customers
- Rewarding to tell a story from data
- Fast moving field, technical results fast changing
- Applicable virtually anywhere
- Increasing amounts of open data available
- Creative activity
- Can be used for social good
  - Pro-bono work for data science students?
  - E.g. https://www.meetup.com/en-AU/The-Minerva-Collective/
  - Healthy salary

Challenging aspects

- Capabilities of technology versus ethics and privacy
  - Cambridge Analytica
  - Google duplex
  - …
- GDPR: European General Data Protection Regulation
  - https://www.cnet.com/how-to/what-the-gdpr-means-for-facebook-the-eu-and-you/

Data science: Types of jobs

- Some companies employing our students in data scientist roles
  - Telstra, Citibank, Danske Bank, Deutsche Bank, NAB, ANZ, Veda, Tencent, LexisNexis Risk Solutions, GE Capital, Deloitte, PwC, Accenture, Deloitte, IBM Research, IBM, Sportsbet, OpenBet:, CrowdsourceHire, Hugo, Flipkart, Rome2rio, Breadtrip, SAP, Salesforce, Hitachi, Oracle, Google Apple, Microsoft, Amazon, Groupon, Nokia, CSIRO, MongoDB, DST Group, Data61, Evernote, Teradata, Kepler Analytics, Business Predictions, Thales, Tata, LinkedIn, Ford, Huawei, KPMG, northraine, Woolworths, jet.com, Microsoft Research, SAS, Peter MacCallum Cancer Centre, Commonwealth Bank, Computershare, Blackmagic Design, Baker IDI, AIG, ….

Exam

- 2 hours
- 50 marks (50% of subject overall mark)
  - Must obtain at least 20/50 marks in exam as hurdle requirement

Past exams

- 2016 exam
  - Ignore questions 1b, 1c, 4c, 4e
- 2017 and 2018-sm1 exams
  - All questions relevant
- Sample exam and sketch answers
  - Available on the LMS
- What's examinable?
  - Refer to exam study guide on the LMS
- First two pages of the 2018 exam (cover sheet and formulas)
  - Available in the exam study guide

Exam

- You don't need to write programs in Python
  - If you are asked to write anything algorithmic, you may use pseudo code
- A number of formulas will be provided on the exam
  - Provided at the end of the study guide
- 17 questions
  - Multiple parts to some questions
  - Start each question (but each part of a question) on a new page
  - Point form answers are encouraged

Typical types of question

- Explain concepts
  - Define how 3 party protocol for privacy preserving operates
- Use concepts
  - Perform an imputation method
  - Compute a similarity
- Understand limitations/scope/importance of concepts
  - Would Pearson correlation correlation coefficient be appropriate for this example ..?
  - Advantages of PCA for a particular example …?
- Compare concepts (advantages/disadvantages/similarities/differences)
    - Compare HTML and XML

Marking

- As a *general* guide
  - If question is worth 3 marks, we will be looking for 3 items in the answer (each 1-3 sentences in length)
  - If question is worth 4 marks, we will be looking for 4 items in the answer (each 1-3 sentences in length)
  - ….
  - Avoid irrelevant or excessively long answers

Revision: User-user similarity

U1	**17**	**-**	**20**	**18**	**17**	**18.5**
U2	**8**	**-**	**-**	**17**	**14**	**17.5**

- Method1:

$$SIM(U1, U2) =$$
$$((17-8)^2 + (18.1-14.1)^2 + (20-14.1)^2 + (18-17)^2 + (17-14)^2 + (18.5-17.5)^2$$

- Method2:

$$Sim(User1, User2) = \frac{6}{6-2}((17-8)^2 + (18-17)^2 + (17-14)^2 + (18.5-17.5)^2$$

- For user-user method2, we consider using squared Euclidean distance

Revision: Item-user similarity (1)

Users	Titanic	Batman	Inception	Jurassic World	Superman
Michelle	2.5		3	3.5	3
Tom	3	3.5		5	3.5
Lao	2.5	3		3.5	4
Jane	?	3.5	3	4	2

- Item j: Titanic
- User a: Jane

Step 1: calculate k-most similar items to item j:
- Let's say k = 3

Distance	Batman	Inception	Jurassic World	Superman
Titanic – Method 2	$\frac{4}{2}(0.5^2 + 0.5^2)$	$\frac{4}{1}(0.5^2)$	$\frac{4}{3}(1+4+1)$	$\frac{4}{3}(0.5^2 + 0.5^2 + 1.5^2)$
Similarity	$1/1$	$1/1$	1/8	$1/\frac{11}{3}$

Revision: Item-user similarity (2)

Users	Titanic	Batman	Inception	Jurassic World	Superman
Michelle	2.5		3	3.5	3
Tom	3	3.5		5	3.5
Lao	2.5	3		3.5	4
Jane	?	3.5	3	4	2

- Item j: Titanic
- User a: Jane

$$r_{aj} = \frac{\sum_{i \in \text{k-similar items}} sim(i,j) \times r_{ai}}{\sum_{\in \text{k-similar items}} sim(i,j)}$$

Step 2: r_{aj}
- In the rating example: similarity = 1/squared Euclidean distance

$$r_{aj} = \frac{1x3.5 + 1x\ 3\ + \frac{3}{11}x\ 2}{1 + 1 + \frac{3}{11}}$$

Revision: Item-user similarity (3)

Update in lecture 5 slide 26

$$r_{aj} = \frac{\frac{1}{3.5}x5 + \frac{1}{1.65}x\ 3.5 + \frac{1}{3.5}x\ 3}{\frac{1}{3.5} + \frac{1}{1.65} + \frac{1}{3.5}}$$

Revision: 1st and 3rd quartile (1)

Finding Q1 and Q3: Method 1

- Step 1: Put the numbers in order: 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27
- Step 2: Find the median: 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Step 3: Find the median of the upper set of numbers above the median. This is the upper quartile (Q3):
  - 1, 2, 5, 6, 7, 9, (12, 15, 18 ,19 ,27).
- Step 4: Find the median of the lower set of numbers below the median. This is the lower quartile (Q1):
  - (1, 2, 5, 6, 7), 9, 12, 15, 18 , 19 ,27).

Revision: 1st and 3rd quartile (2)

Finding Q1 and Q3: Method 2

- $\text{median} n = \frac{n+1}{2}$, $Q1 = \frac{n+1}{4}$, $Q3 = \frac{3(n+1)}{4}$ [gives the number of the element]
- Step 1: Put the numbers in order: 1, 2, 5, 6, 7, 9, 12, 15, 18, 19, 27.
- Step 2: n = 11
  - Q1= 3rd element = 5
  - Q3= 9th element = 18
  - median = 6th element = 9

Exam 2018

Questions before the exam

- Post in the “Exam discussion” forum on the LMS
- Exam consultation sessions:
  - Monday 22/10/2018 Room 07.02 Doug McDonell 10:00am-12:00pm
  - Thursday 25/10/2018 Room 07.02 Doug McDonell 10:00am-12:00pm

We would like to get your feedback

- The Subject Experience Survey (SES) is open on-line to students until **October 28**
  - https://ses.unimelb.edu.au
- Results of the SES are taken seriously and used to
  - Evaluate teaching staff
  - Improve the subject
- Please do participate!

Final words

- Thanks for participating in the *Elements of Data Processing!*
- The COMP20008 team
  - James, Yasmeen, Anam, Ilya, Sobia and Namrata
- Good luck with your exams and future studies