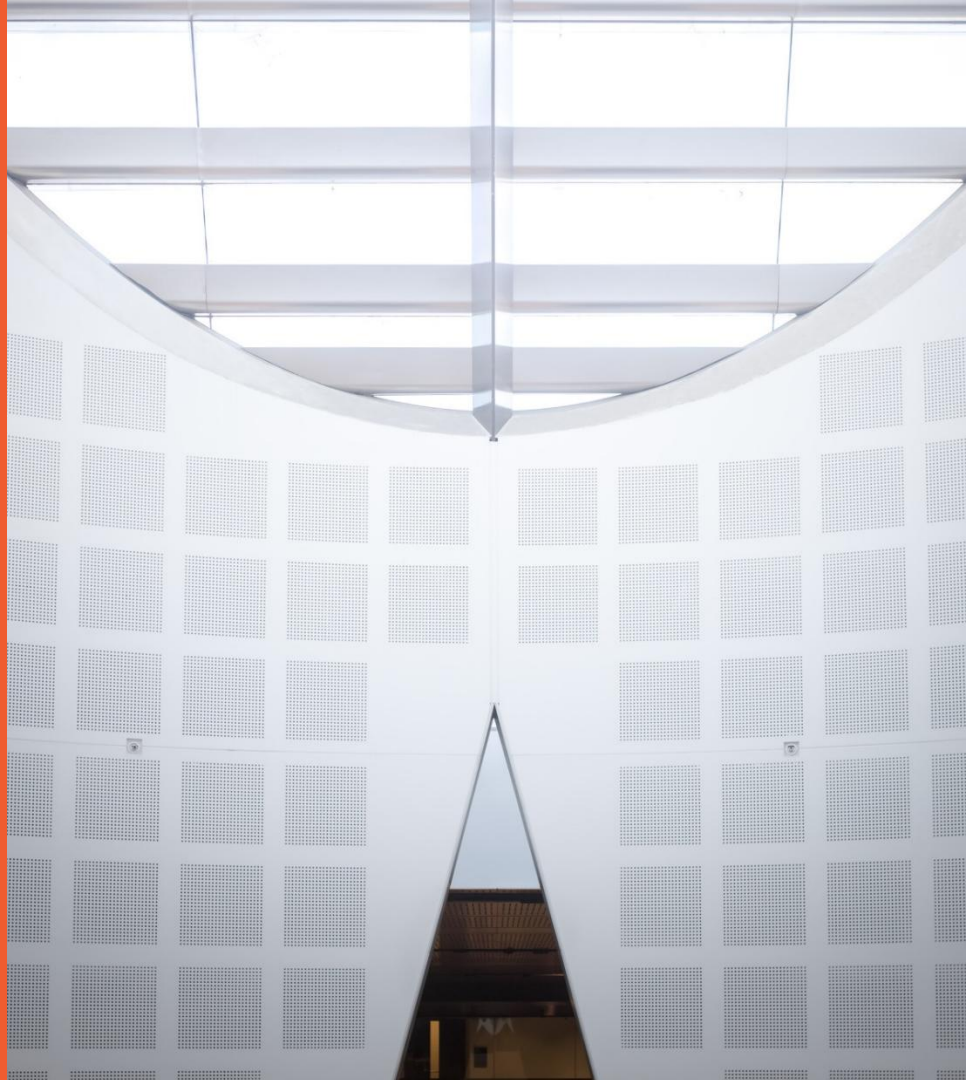
COMP5310: Principles of Data Science

Review

Presented by
Maryam Khanian
School of Computer Science

THE UNIVERSITY OF SYDNEY



Unit of Study Survey (USS)

- If you haven't done so, please spend a few minutes to complete the survey.
  - Scan the QR code below, or browse to: https://student-surveys.sydney.edu.au/students/
  - Log in if you are not already
  - Complete survey for COMP5310

Each survey completed will give an entry into the prize draw for a range of JB HiFi Giftcards totalling $2500

EXAM ARRANGEMENTS

THE UNIVERSITY OF SYDNEY

Assessments

- 15%: Project stage 1
- 25%: Project stage 2
- 60%: Final exam

Exam Schedule

- Duration: 2 hours examination + 10 minutes reading time
- Please check your exam timetable for the venue, day, time
  - On-campus, paper-based exam
  - In-person: by exam supervisors in the room

Exam Condition

- Restricted open book
- Materials permitted
  - Calculator – non-programmable (Calculators with advanced programmable capabilities or artificial intelligence features are strictly prohibited.)
  - **One A4** sheet of handwritten and/or typed notes single-sided

- No need blank paper, since blank pages are already included at the end of the exam paper –use them if needed.

- No other materials are allowed – no lecture slides, tutorial notes, books or other materials
- No other devices are permitted/ No internet browsing is allowed
- You can't consult other people during the exam

In-person exam

- There are specific rules about the in-person exams, read them carefully:
  - https://www.sydney.edu.au/students/exams/in-person.html
- Arrive early – you need enough time to find the exam room and confirm your seat number
- You will not be admitted to the exam more than 30 minutes after your exam has commenced
- You need to have a photo identification – student ID card or another accepted identification
- The exam paper is like a booklet. You write your answers on the exam paper, in the space provided. There is a question, then space for you to write the answer; another question and space for you to answer, etc.
- To write the answers you should use black or blue pen, not pencil

Exam paper is confidential

- The exam paper is confidential
- You must not discuss the exam questions with other people, post or distribute them in any way during the exam or after the exam

Outline of Lectures

Week	Topic
~~Week 1~~	~~Introduction~~
Week 2	Data Cleaning and Exploration (via Spreadsheet)
Week 3	Data Exploration with Python
Week 4	Data Transformation and Storage with Python and SQL
Week 5	Querying and Summarising Data with SQL
Week 6	Hypothesis Testing and Evaluation
Week 7	Association Rule Mining
Week 8	Clustering and Dimensionality Reduction
Week 9	Linear Regression & Logistic Regression
Week 10	Decision Tree
Week 11	Naïve Bayes Classifier
~~Week 12~~	~~Product Thinking & Ethics~~
Week 13	Unit of Study Review

Exam Questions

- No python programming

- Pass the unit of study
  - You must obtain at least 40% (i.e., 40 marks) in the final exam, as well as an overall mark of at least 50 marks, to pass the unit

Exam Advice

- Plan how you will allocate time (wisely)
  - Use “reading time”
  - If you are uncertain about a question during the exam, answer to the best of your ability
- Write clearly
- If you need more space, use blank pages at the end of the exam booklet

Exam Advice

- Find the room location before the exam day!
- Bring a Calculator- non-programmable
- Bring spare pens (either black or blue)
- Have your student ID and put it on the desk

Types of questions

- The exam contains 3 types of questions:
  - Multiple choice questions and T/F
  - Questions requiring short answers / test your understanding and ability to relate concepts; be clear and concise
  - Problem solving/calculation questions

Multiple choice questions OR T/F

In logistic regression, what happens to the cost when the true label is 'Y=1' but the predicted probability $h_\theta(x)$ is close to 0.

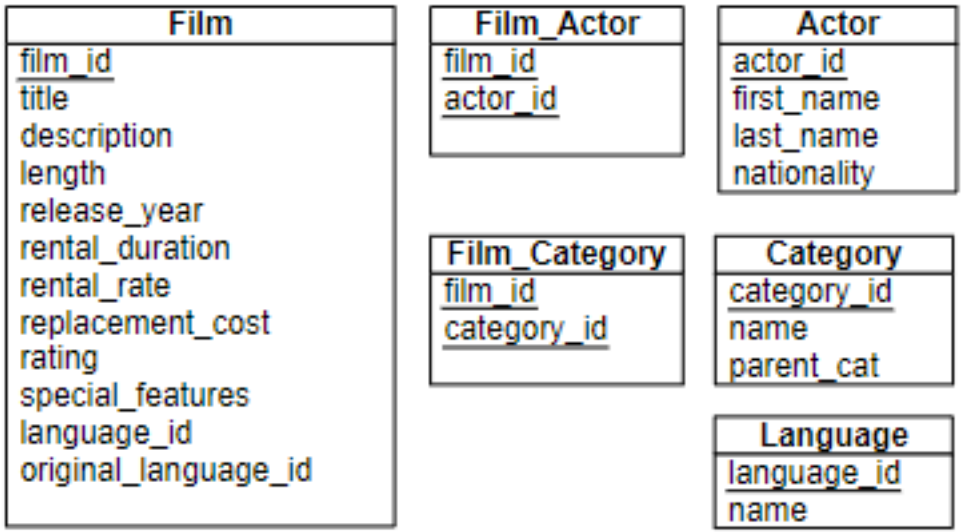
a) The cost is very low.
b) The cost approaches 0.
c) The cost is very high.
d) The cost is undefined.

If there is a very strong correlation between two variables, then the correlation coefficient must be:

a) Any value larger than 1
b) Much smaller than 0, if the correlation is negative
c) Close to +1 or -1, depending on whether the correlation is positive or negative

Short answer questions/ Decision-Making Questions

Consider the following relational schema tables. Write an SQL query that lists the title of every Film in which the Actor named JOHNNY CAGE has appeared. The titles should be in alphabetical order. Use only the necessary tables to extract the required information.



Question: For the following attributes:

Final exam score of Data Science class

- Classify as binary, discrete or continuous.
- Classify as nominal, ordinal, interval or ratio.

Question: How do the mean, median, and mode compare in a normal distribution versus a right-skewed distribution.

Question: If you have a multiple regression model for house prices with coefficients:

$\beta_0 = 50{,}000$

$\beta_1 = 100$ (for square footage)

$\beta_2 = 5{,}000$ (for number of bedrooms)

Write out the full equation of the model.

Problem solving/calculation questions

Consider the training examples shown in the Table below for a binary classification problem.

We would like to build a decision tree using information gain. Which attribute will be selected as a root of the tree? Show your calculations.

Training examples: 9 True/ 5 False

Class label

Applicant	Level	Lang	Tweets	PhD	Interviewed well
A1	Senior	Java	No	No	False
A2	Senior	Java	No	Yes	False
A3	Mid	Java	No	No	True
A4	Junior	Python	No	No	True
A5	Junior	R	Yes	No	True
A6	Junior	R	Yes	Yes	False
A7	Mid	R	Yes	Yes	True
A8	Senior	Python	No	No	False
A9	Senior	R	Yes	No	True
A10	Junior	Python	Yes	No	True
A11	Senior	Python	Yes	Yes	True
A12	Mid	Python	No	Yes	True
A13	Mid	Java	Yes	No	True
A14	Junior	Python	No	Yes	False

Calculation

- $H(\text{Interviewed well}) = H(9,5) = -({}^{9}/_{14}\log_2({}^{9}/_{14}) + {}^{5}/_{14}\log_2({}^{5}/_{14})) = 0.94$
- $H(\text{Interviewed well} \mid \text{Level}) = \sum_i p(\text{Level} = v_i) * H(\text{Interviewed well} \mid \text{Level} = v_i)$

v_i	$p(\text{Level} = v_i)$	$H(\text{Interviewed well} \mid \text{Level} = v_i)$
Senior	${}^{5}/_{14} = 0.36$	$H(2,3) = -\left({}^{2}/_{5} * \log_2\left({}^{2}/_{5}\right) + {}^{3}/_{5} * \log_2\left({}^{3}/_{5}\right)\right) = 0.97$
Mid	${}^{4}/_{14} = 0.29$	$H(4,0) = -({}^{4}/_{4} * \log_2({}^{4}/_{4}) + {}^{0}/_{4} * \log_2({}^{0}/_{4})) = 0$
Junior	${}^{5}/_{14} = 0.36$	$H(3,2) = -\left({}^{3}/_{5} * \log_2\left({}^{3}/_{5}\right) + {}^{2}/_{5} * \log_2\left({}^{2}/_{5}\right)\right) = 0.97$

- Then:
  - $H(\text{Interviewed well} \mid \text{Level}) = 0.36 * 0.97 + 0.29 * 0 + 0.3 * 0.97 = 0.7$
- Thus:
  - $IG(\text{Interviewed well} \mid \text{Level}) = H(\text{Interviewed well}) - H(\text{Interviewed well} \mid \text{Level}) = 0.94 - 0.7 = 0.24$

Calculation

- Similarly:
  - $IG(\text{Interviewed well} \mid \text{Tweets}) = H(\text{Interviewed well}) - H(\text{Interviewed well} \mid \text{Tweets}) = 0.15$
  - $IG(\text{Interviewed well} \mid \text{PhD}) = H(\text{Interviewed well}) - H(\text{Interviewed well} \mid \text{PhD}) = 0.048$
  - $IG(\text{Interviewed well} \mid \text{Lang}) = H(\text{Interviewed well}) - H(\text{Interviewed well} \mid \text{Lang}) = 0.029$
- Level has the highest information gain, therefore it was good to choose that attribute.

A national park has created a dataset to help hikers determine if a reptile they encounter could be venomous.

	Head	Eyes	Size	Venomous
1	Triangle	Elliptical	Small	Yes
2	Round	Round	Small	No
3	Narrow	Elliptical	Small	No
4	Narrow	Round	Large	No
5	Narrow	Elliptical	Large	Yes
6	Triangle	Round	Small	Yes
7	Narrow	Round	Large	No
8	Round	Elliptical	Large	No
9	Triangle	Elliptical	Small	Yes

Use Naïve Bayes to predict if the following example is venomous or not:

Head=narrow, Eyes=elliptical, Size=Large

Show the working for your calculations.

- We hope you found this course **useful**!
- Good luck with the exam and best wishes from the whole teaching team!