COMP9318: Data Warehousing and Data Mining

Course Introduction

What is Data Warehousing?

- "A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process." — W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses
- Difference between data warehouse and database

What is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Alternative names
    - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- In this course, we will cover several major topics in data mining
  - Classification
  - Clustering
  - Association rule mining
  - …

Lecture in Charge

- Lecturer-in-charge:
  - Dr. Yifang Sun
  - School of Computer Science and Engineering
  - office: K17-208
  - email: yifangs@cse.unsw.edu.au
    - use [comp9318] in subject

- Research interests
  - High dimensional data
  - Machine learning (Natural language processing)
  - Knowledge graph
  - Integration of DB and AI

Course Aims

- Introduce the the foundation of data warehousing
  - OLAP
- Introduce the theories of various data mining techniques
  - Classification
  - Clustering
  - Association rules
  - …
- Explore the practice of developing data mining applications
  - Programming project
  - Labs
  - …

Course Aims - cont.

- Not possible to cover every aspect of data warehousing and data mining
- We will focus on
  - concepts
  - algorithms
  - principles
- We will not focus on
  - programming languages and API
  - specific platforms/tools
- Make use of tutorials and documents on the Internet

Lectures

- Delivered through pre-recorded videos
  - location: anywhere you like
  - time: anytime you want
  - links to videos available on Ed before the lecture
    - email LiC ASAP if you have problem access to Ed
- Slides on course website
- No QA sessions during lectures
  - Ask in the forum or during online consultations
  - Will address common questions at the beginning of each lecture
- Schedule and length of lectures may vary based on the progress of the course
- Note: watching every lecture is assumed

Consultations

- Online QA discussions using Ed
  - encourage you all to participant
  - Raise questions and try to help others
- Online consultation with tutor
  - 12pm – 1pm every Friday
  - using Zoom
  - room number and password will be in Ed
- Private online consultation with LiC
  - please book an appointment with me with a brief description of your questions, with [comp9318] in subject
  - only for problems cannot solve in the forum and during the online consultation

Resources

- Lecture Slides
  - Contains many materials not found in the text/reference books.
- Textbooks
  - Jensen et al, Multidimensional Databases and Data Warehousing. (Accessible from a UNSW IP)
  - Han et al, Data Mining: Concepts and Techniques, 1st/2nd edition, Kaufmann Publishers.
- Reference Books
  - Charu Aggarwal, Data Mining: The Textbook, Springer, 2015.
  - Tan et al, Introduction to Data Mining, Addison-Wesley, 2005.
  - Leskovec et al, Mining of Massive Datasets (ver 2.1), Available at http://infolab.stanford.edu/~ullman/mmds.html

Resources - cont.

- Software
  - Anaconda
  - Python 3
  - Jupyter notebook
  - Python libs such as numpy, pandas, matplotlib, scikit-learn, . . .
- Reading Materials
  - Papers from machine learning/data mining conferences/journals, white papers, surveys, etc.
  - All available from the course Web page
- Online Resources
  - Online courses and tutorials from YouTube, Coursera

Pre-requisite

- Official prerequisites
  - Discrete mathematics (COMP9020)
  - Data Structures and Algorithms (COMP9024)
  - Database Systems (COMP9311)
- Before commencing this course, you should
  - have experiences and good knowledge of algorithm design
  - have solid background in database systems
  - have solid programming skills in Python
  - be familiar with Linux operating systems
  - have basic knowledge of linear algebra, probability theory and statistics

Please do not enrol if you…

- Don't have COMP9020/9024/9311 knowledge
- Cannot produce correct Python program on your own
- Have poor time management
- Are too busy to watch lecture videos/labs

- Otherwise, you are likely to perform badly in this subject

Assessment

- Five labs (full mark: 25)
  - Only the best 3 will be counted
- One written assignment (full mark: 25)
- One programming project (full mark: 50)
- Final exam (full mark: 100)
  - Double pass (>=40)

- Final Mark $= \frac{2 \cdot (labs + assn + proj) \cdot final}{labs + assn + proj + final}$ if $final \geq 40$
  - FL if $final < 40$

Labs

- Labs to help you with programming and projects
- Only the best 3 will be counted
  - lab = np.mean(sorted([lab1, lab2, lab3, lab4, lab5], reverse=True)[:3])
- Unlimited trials
- Immediate feedback
  - Don’t rely on the feedback and blindly try
- No late submission allowed for labs

Written Assignment

- Exam-style questions
  - Computational, short answer
  - no essay, no multiple choice
- Regarding the lecture contents
  - algorithms, principles, …
  - to assess your understanding, not memory
- Late penalty
  - firm deadline
  - zero mark for late submission

Programming project

- Individual task
- Both results and source codes will be checked.
  - Zero mark if your codes cannot be run due to some bugs.
- Late penalty
  - 10% reduction of raw marks for the 1st day, 30% reduction per day for the following 3 days

Final exam

- Open book exam
- Firm deadline
- No supplementary exam will be given if you fail
- Special consideration must be submitted prior to the start of the exam

- More details on the way

Tentative course schedule

Week	Topic	Labs/Assignment/Project
1	Course Introduction and Math review	
2	Data warehousing and OLAP	lab1
3	Data preprocessing	
4	Classification	lab2
5	Classification	lab3
6	Flexibility Week (no lecture)	project
7	Clustering	Assignment
8	Clustering	lab4
9	Association rule mining	
10	Revision and Exam Preparation	lab5

Warning

- This course has
  - Broad coverage
  - Heavy workload
  - High fail rate $\geq 20\%$
- Specially, we do not accept personal plea or excuses
  - if you have valid reasons that affect your performance, apply for a UNSW Special Consideration
    - https://student.unsw.edu.au/special-consideration.

Warning - cont.

- Common excuses
  - I spent so much time and effort on this course but still failed?
  - I did the work by myself and may have shared it with my classmate for discussion.
  - If I fail this course, I will [...]. Please.

- We aim to build a fair environment for every student in this course

Academic honesty and plagiarism

- Zero tolerance to plagiarism
  - You will get 0 marks
- Examples of misconduct:
  - Copy other students' work
  - Let other students copy your work
  - Copy from GitHub
  - Find a ghost writer
  - …
- I will not accept the following excuses:
  - "I've left the lab with my screen unlocked"
  - "He stole it from my computer"
  - "I only gave my code to A. A didn't use it but gave it to B"
  - …
- Make sure you read all types of plagiarism, esp. collusion in https://student.unsw.edu.au/plagiarism.

General Recommendations

- Make use of LiC and tutors
  - don't hesitate to ask questions
- Make use of the forum
  - read the notices in course website and Ed
  - participate in the discussions in Ed
- Make use of course materials
  - understand lecture slides
  - read specifications carefully
  - try all the labs although they are not compulsory
- Do not misconduct

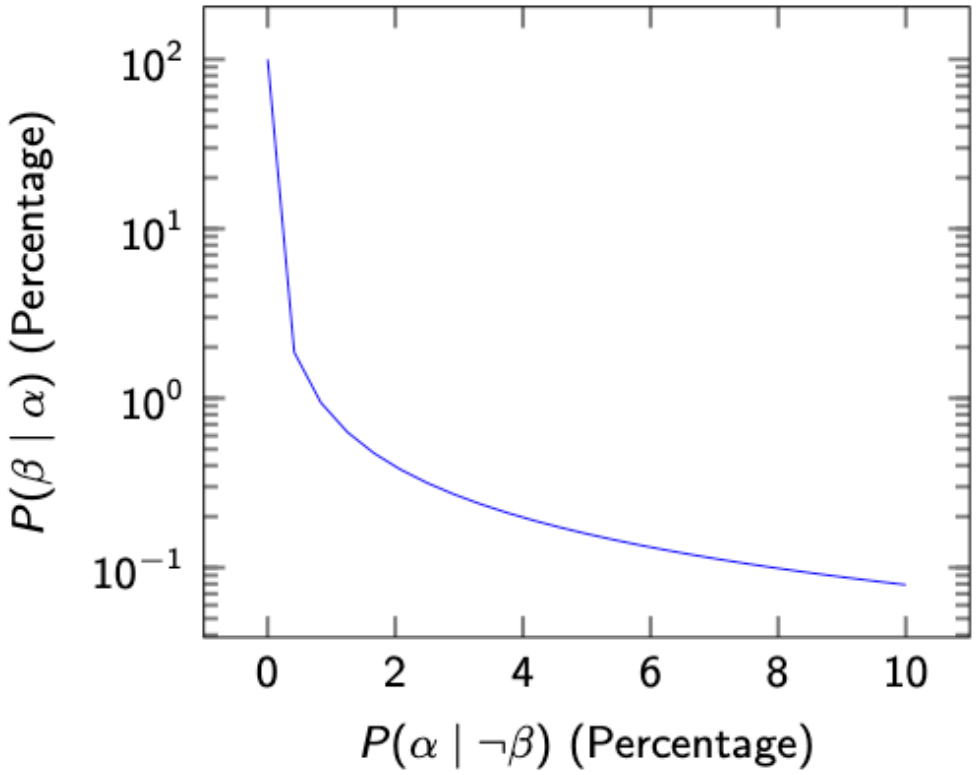
About Learning

- Understand (not memorize) concepts/equations/algorithms
  - Ask why
  - Describe it in you own language to a layman

- Example
  - Phil got a positive result for the α test and the probability that patients with the deadly ß disease having a positive α test is 99%. Should Phil be worried about having the ß disease?

- Example
  - Phil got a positive result for the α test and the probability that patients with the deadly ß disease having a positive α test is 99%. Should Phil be worried about having the ß disease?

- Plot the function Pr[ß|α] with respect to Pr[α|not ß] given Pr[ß]=0.00008



- Example
  - Phil got a positive result for the α test.
  - All patients with the deadly ß disease have a positive α test result.
  - Does Phil have the ß disease?

Your Feedbacks are Always Welcome

- Please advice where I can improve after each lecture, through Ed or by email

- myExperience system