About COMP9318 (2019 t1)

Wei Wang @ CSE, UNSW

February 17, 2019

Introduction

Lecturer-in-charge:

Prof. Wei Wang

School of Computer Science and Engineering
Office: K17 507
E-mail: weiw@cse
Ext: 9385 7162
http://www.cse.unsw.edu.au/~weiw

Research Interests:

- Knowledge graph / natural language processing
- High-dimensional data / Similarity query processing
- AI security
- DB + AI

Course Info

- Homepage: `http://www.cse.unsw.edu.au/~cs9318`
- Communications:
  - Main form: Piazza Forum: TBA on the course web site
  - **Email**: `weiw AT cse.unsw.edu.au`:
    - Only for matters that cannot/should not be resolved via piazza.
- Lectures:
  - 1800 – 2100 MON, Keith Burrows Theatre
- Tutorials: several *online* tutorials + `ipython notebooks`
- Consultations:
  - Weekly by tutors (TBA)
  - LiC: in lectures or by appointment only.

Assessment

Overview

- 1 written assignments + 1 programming project + lab
- `lab = np.mean(sorted([lab1, lab2, lab3, lab4, lab5], reverse=True)[:3])`

No late submission allowed for labs. Read the spec of assignment/project to find out late penalty policies.

Project

Default project

- up to 2 students per team
- Hidden Markov Model

Research project

- Apply by Week 3
- Any research topic related to Data Mining, Machine Learning/Deep Learning, Natural Language Processing, Interpretability, Adversarial Machine Learning.
- Will post some suggested papers/links on the course page

Finally . . .

Exam

- If you are ill on the day of the exam, **do not attend** the exam — I will **not** accept medical special consideration claims from people who have already attempted the exam.

Final Mark

- Final mark

  $$final\_mark = 0.15 \cdot ass1 + 0.20 \cdot proj1 + 0.10 \cdot lab + 0.55 \cdot exam$$

- Also requires $exam \geq 40$.

Warning I

This course has

- Broad coverage
- Heavy workload
- High fail rate $\geq 20\%$
- Plagiarism is not allowed. Make sure you read all *types* of plagiarism, esp. collusion in `https://student.unsw.edu.au/plagiarism`.

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 II

Example excuse

- 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.

Resources I

Lecture Slides

- Contains many materials not found in the text/reference books.

Text Book

- Leskovec *et al*, *Mining of Massive Datasets* (ver 2.1), Available at `http://infolab.stanford.edu/~ullman/mmds.html`
- 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

- Tan *et al*, *Introduction to Data Mining*, Addison-Wesley, 2005.

Resources II

- Witten *et al*, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 1st/2nd edition, Morgan Kaufmann.
- Charu Aggarwal, *Data Mining: The Textbook*, Springer, 2015.

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.

Schedule (tentative)

Week	Contents	Assignments
1	Course overview, Introduction, Data warehousing and OLAP	
2	Data warehousing and OLAP + Maths review	lab1
3	Data Preprocessing	
4	Classification	lab2
5	Classification	assignment
6	Classification	lab3
7	Hidden Markov Model	project
8	Clustering	lab4
9	Clustering	
10	Association Rule Mining + (misc)	lab5

Course Objective and Requirements

Objectives:

- Cover practically useful data mining/machine learning algorithms and concepts
- Foster **deeper** understanding of maths, models, and algorithms
- Gain hands-on experience with solving real problems

Requirements:

- You need to have a solid background in Maths (Linear Algebra, Calculus, Probability & Statistics) and programming (mainly python).
- Understand (not memorize) concepts/equations/algorithms.
  - Ask *why*.
  - Describe it in your own language to a layman.

Feedback welcome (throughout the course).

Example

Example

John got a positive result for the α test, and the probability that patients with the deadly β disease having a positive α test result is 99%. Should John be worried about having the β disease?

Example

Example

John got a positive result for the α test, and the probability that patients with the deadly β disease having a positive α test result is 99%. Should John be worried about having the β disease?

$$P(\beta \mid \alpha) = \frac{P(\alpha \mid \beta)P(\beta)}{P(\alpha)} = 0.99\frac{P(\beta)}{P(\alpha)}$$

Example

Example

John got a positive result for the α test, and the probability that patients with the deadly β disease having a positive α test result is 99%. Should John be worried about having the β disease?

$$P(\beta \mid \alpha) = \frac{P(\alpha \mid \beta)P(\beta)}{P(\alpha)} = 0.99\frac{P(\beta)}{P(\alpha)}$$

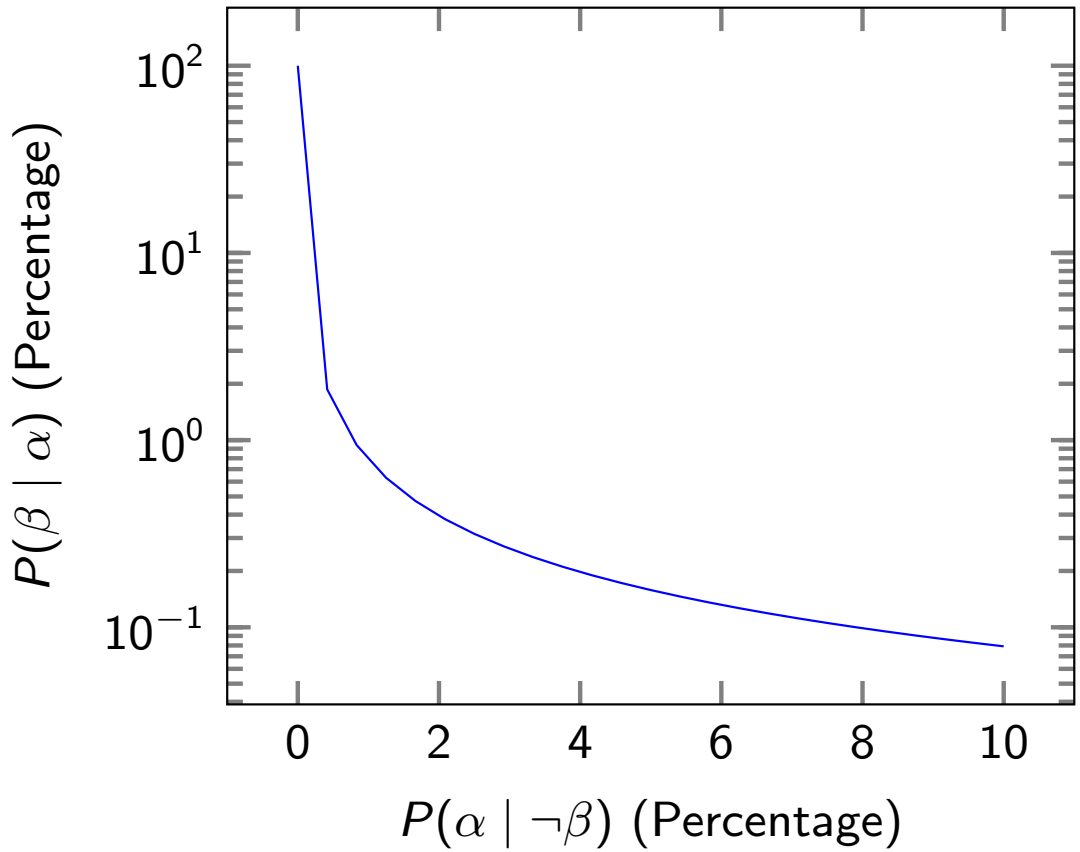
$$P(\beta \mid \alpha) = \frac{P(\alpha \mid \beta)P(\beta)}{P(\alpha \mid \beta)P(\beta) + P(\alpha \mid \neg\beta)P(\neg\beta)}$$

false postive rate

Example

Exercise

Exercise: plot the function $P(\beta \mid \alpha)$ with respect to $P(\alpha \mid \neg\beta)$ given $P(\beta) = \frac{8}{100{,}000}$.



CSE Computing Environment

For those new to the computing environment at CSE, UNSW

- Use Linux/command line.
  - Project marked on linux servers
  - You need to be able to upload, run, and test your program under linux.
- Assignment/Project submission
  - `Give` to submit. Watch out for possible error messages.
  - `Classrun`. Check your submission, marks, etc. Read `https://wiki.cse.unsw.edu.au/give/Classrun`
  - Common errors:
    - File corrupt (during SFTP?), not in the correct format.
    - Submission not accepted by the system (wrong filename? too large? ...).
- Lab submission: our home-made Web submission system.

Other Specialised Courses

Other specialised courses in the Database or Data Science stream:

- COMP9319: Advanced algorithms on compression, text/XML databases, etc.
- COMP9313: Big data systems (hadoop, spark, etc)
- COMP6714: Information retrieval, Natural language processing, Search engines.

Other machine learning courses:

- COMP9417: Machine Learning and Data Mining
- COMP9444: Neural Networks and Deep Learning
- COMP9418: Advanced Machine Learning

About Learning

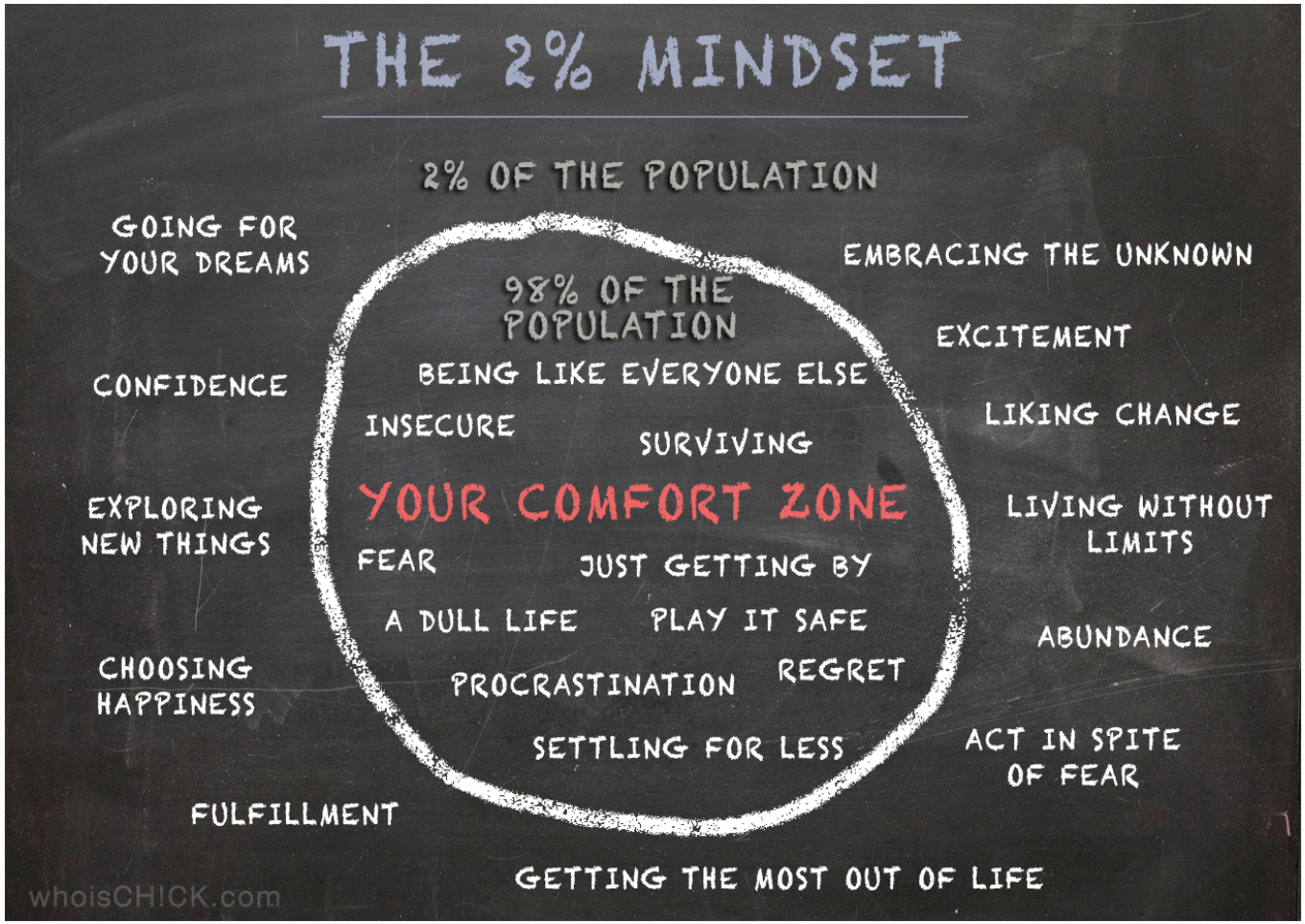
Things to ponder:

- The long-term impact of the latest development in AI/DS/Hardware.
- What do you want out of this course?

Requirement:

- Plan ahead for the course.
- Learning happens outside your comfortable zone.
- Review teaching materials after the lecture.
- Use the Jupyter notebooks.

Make Errors and Learning Sth. New



Source:

`http://combiboilersleeds.com/images/comfort-zone/comfort-zone-0.jpg`