BMEG 3105: Data Analytics for Personalized Genomics and Precision Medicine

Fall 2024

Lecture 1 – Course Introduction

Lecturer: Professor Yu Li

Scribe: Lau Kwan Yee

Outline of lecture 1

A) Review of Pre-course Survey Results
B) Course Logistics
C) Course Grading System
D) Brief Overview of DATA in Personalized Genomics and Precision Medicine
E) Expected Outcome

A. Review of Pre-course Survey Results

- Response: 47/52
- Student's Background:

  39 students from BME, 3 students from CS, 3 students from AIST, 1 student from Cell and Molecular Biology and 1 student from AI

  42.6% from Year 3, 40.4% from Year 4, 10.6% from Year 2 and the rest from Year 5

Q1. Major drivers of taking this course: (with the 3 highest average score)

1) For the three credits and degree requirement
2) For biological/ genomics/ health applications
3) Just exploring a new field

Q2. Areas that students are not familiar with: (with the 3 lowest average score)

1) Machine learning
2) Algorithms
3) Programming; Research

Q3. Top 5 topics students are interested in:

1) Protein-protein/ RNA interaction
2) Neutral networks
3) Dynamic programming
4) Convolutional neural networks
5) Cancer genomics

Q4. Top 3 problems students need assistance with:

1) Implementation the methods with programming
2) Apply the methods learned from the course to new data
3) Understand the concept in data analytics

B. Course Logistics

- Timeslots:

Lectures: Wed 9:30am – 11:15am (11:05am) SC L4
Fri 9:30am – 10:15am MMW 703
Tutorials: Fri 10:30am – 11:15am MMW 703
*Lectures will be video recorded and available after the lecture
* TA will have several sessions to help with Python programming (exact schedule determined by the TA)

Course website: https://yu979.github.io/BMEG3105-Fall-2024/
*Slides will be available on the day before the lecture

- Teaching team:

1. Instructor: Yu Li
   Email: liyu@cse.cuhk.edu.hk
   Office hour: Fri 3pm – 5pm
   Location: SHB 106 (or on request)
2. TA: Qinze Yu
   Email: qzyu22@cse.cuhk.edu.hk
   Office hour: Mon 2pm – 4pm
   Location: SHB 116
3. TA: Yimin Fan

Email: fanyimin@link.cuhk.edu.hk

Office hour: Tue 2pm – 4pm

Location: SHB 904

4. TA: Ziqian Lin

Email: linziqian@link.cuhk.edu.hk

Office hour: Fri 3pm – 5pm

Location: SHB 904

- Software and communications

Blackboard: the main software to manage the course and grading will be through Blackboard

Piazza [http://piazza.com/cuhk.edu.hk/fall2024/bmeg3105] : To ask questions

*Please use the private post to the instructor or TA for personal matters

C. Course Grading System

1. Homework (20%) – 3 grading homework (5%+5%+5%) + 1 non-grading programming assignment (5%)
    1. Programming Assignment 0: Programming environment setup
        Posted: Sep 6 Due: Sep 20 (no need to submit anything)
    2. Assignment 1: About the basic concept of data analytics – 1
        Posted: Sep 13 Due: Sep 27
    3. Assignment 2: About the basic concept of data analytics – 2
        Posted: Oct 4 Due: Oct 18
    4. Programming Assignment 1: About Application of DA to the biology
        Posted: Oct 30 Due: Nov 15
    5. Assignment 3: DA in personalized genomics and precision medicine
        Posted: Nov 13 Due: Nov 22
2. Scribing (10%)

    - Summarize one of the lectures

    - Submit it within one week after the lecture

    *First two lectures have additional one week for submission

    *Grades will be deducted by 25% for each additional late day

    - Each student should do at least one lecture and the notes will be posted online for others reference

*You can sign for at most two for additional 1%

3. In-class quiz (10%)

   - Dates: Oct 18 & Nov 27

   - The questions will be simple, and the quizzes will be open book. It is mainly for checking the participation

4. Midterm (20%)

   - Date: Oct 23

   - Open book (only paper-based materials are allowed)

   - One bonus question (2%)

   - To know how you understand the concept of data analytics and thus adjust the left half semester

5. Final (20%)

   - Open book (only paper-based materials are allowed)

6. Project (20%)

   - Individual project

   - Potential projects: bio-image classification, gene enrichment analysis, etc.

i. Project milestone report (Proposal) (5%)

   Due: Nov 8

   1 page report, clarifying the following things:

   - Title, author
   - What problem do you want to do? Why is the problem interesting? (1%)
   - What data are you going to process? The source, size and sample of the data (1%)
   - What's the output of your method? (1%)
   - How are you going to do it? Describe the method step by step, from input to output (1%)
   - What are the expected results? How are you going to evaluate the results? (1%)
   - What have you done?

ii. Project Report (7%)

   Due: Dec 2

*Should be submitted with codes (5%) whether it is correct or not

There is no length requirement, the following things should be mentioned:

- Title, author
- What problem do you want to do? Why is the problem interesting and important? (0.5%)
- What data are you going to processed? The source, size and sample of the data (0.5%)
- What have you done to resolve the problem? Describe the method step by step, from input to output (2%)
- What are the results? (1.5%)
- Result evaluation (1.5%)
- Any idea of further improvement? (1%)

iii. Project Presentation (3%)

Date: Nov 22 & Nov 29

7min for each student

The presentation will be evaluated in the following way:

- Logic (1%)
  - What is the problem?
  - Why is it important?
  - How do you resolve it? The overview of your idea
  - The overview of the results
- Clarity (1%)
  - Whether the audience can understand and follow the presentation
- Slides preparation (1%)
  - Clear illustration
  - No typo, no grammar error

7. Bonus (Up to 6%)

One bonus question in Midterm (2%)

One additional scribing (1%)

Pre-course survey + Post-course survey (0.5% each, and the maximum is 3%)

*Fill in the excel form after completing the surveys
https://docs.google.com/spreadsheets/d/1xYWAb7NcAQWl1i3lEX4McRZ_bvnWel3f PE409UXp2Xw/edit?usp=sharing

8. Late Days
   Can be used on A1, A2, A3, PA1, project mid-term report, but cannot be used on final project report and scribing report
   6 late days in total, 2 max for any assignment
   Grades will be deducted by 25% for each additional late day
   *Let the TA know when you want to use the late days

D. Brief Overview of DATA in Personalized Genomics and Precision Medicine

- Reasons for utilizing data analytics:
  - Massive amount of data is being collected and warehoused (E.g. Web, Biological, Bank/ credit card transaction or mobile data)
  - Computers (tools for completing data analytics) have become cheaper and more powerful
  - Data analytics are useful for aggregating data, generating hypothesis, and supporting the conclusion
- Reason for utilizing data analytics in personalized genomics and precision medicine:
  - Sequencing cost decreasing dramatically
  - Single cell data accumulating
  - Global efforts in building biobank
    → lots of sequencing and health data available and waiting to be analyzed
- Data that can be used to measure a person (from inner factor to outer factor)
  - Gene and mutations
  - Gene expression (transcriptome)
  - Proteome
  - Metabolome
  - Molecular network & cellular network
  - Microbiome (oral and gut)
  - Organ (biomedical imaging)

- Hospital test (blood test and so on)
- Electrocardiography (ECG)
- Demographic information (age, gender, location and so on)
- Drug history and disease history
- Personal statement and doctor diagnosis
- Living habit (exercise)
- Diet
- Family history
- Communications and social media data
- Environment (pollution)
- Travel history (global pandemic)

E. Expected Outcome

- Learn the fundamental concept of data analytics
- Know the various data in genomics and medicine
- Apply the data analytics techniques to process the data and resolve problems in biology