BIOENG-210: Biological Data Science I: Statistical Learning

BIOENG-210 Guide to the Course
Prof. Gioele La Manno

February 2024

1 Why BIOENG-210?

Modern biology increasingly relies on our ability to process, analyze, and interpret large datasets. This course provides you with the theoretical foundations, analytical techniques, and software tools necessary to effectively manage and derive insights from complex biological data. Particular focus will be give to the ability to reason about the data and analyze it using the theoretical knowledge you will acquire during the course.

2 Learning Outcomes

By the end of the course, you will be able to analyze multidimensional biological data, apply and interpret various statistical models, and plan end-to-end data analysis pipelines. Most importantly, you will develop the critical thinking skills needed to choose appropriate analytical methods for specific biological problems and interpret their results meaningfully. A complete **syllabus** will be made available on the course website later during the course.

3 Course Structure and Teaching Approach

The course combines theoretical lectures with hands-on **exercises** each week.

3.1 Lectures

The lectures will be delivered through **a slide deck**, which will be made **available** *one week before the lecture* for you to have a preview and to take notes during class. The day of the lecture we will also release detailed **lecture notes**. These lecture notes are intended as the main reference material for the course, in other words, the alternative to a **textbook**.

3.2 Supplementary Reading

For further reading, for lecture we will suggest some **supplementary reading** chapters from different books: Therefore these readings are not mandatory but they will help to deepen the understanding of the topics. The reference textbooks are:

- *ITB - Introduction to Probability* by Joseph K. Blitzstein and Jessica Hwang
- *CASI - Computer Age Statistical Inference* by Bradley Efron and Trevor Hastie
- *ESL - The Elements of Statistical Learning* by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
- *PRML - Pattern Recognition and Machine Learning* by Christopher M. Bishop

Please note the **exam** will be as much as possible limited to the material lecture notes, what is discussed in class and exercise material. Obviously this being a course focused on reasoning and analysis, questions may require applying concepts to new situations.

3.3 Exercises / Hands-on Sessions

In this course, exercises are crucial for developing your data science skills. To emphasize their importance, these exercises are graded (see below).

For the **exercise** sessions, we have adopted a *partially flipped classroom* approach. You will receive access to the **Hands-on** material *the day of the corresponding lecture*, giving time to familiarize with the concepts, prepare questions or even do the exercise before showing up. The exercise material will have a form of a **Jupyter notebook** with analyses tasks to complete. The exercises are in *Python*. They will be released the same day of the lecture of the same week so you can get started if you want.

These **Hands-on** are designed to be engaging and challenging, offering real-world applications of the concepts covered in lectures. Since we are simulating a real world analysis scenario **you are allowed to use AI tools** to help you with your coding, but you should lead the analysis task and be able to understand and explain the code generated. The use of AI to solve the entirety of the exercise, using our material as a prompt is **not allowed** and will be considered cheating. To not impair learning we do not recommend using the smartest models such as "o3" and the "Chat" version of these tools, which will tend to provide complete answer. Instead we recommend the use of VSCode + GithubCopilot with a model such as "4o". In this way you can get inline support to fix bugs, to suggest you the API calls of the libraries you do not yet.

During the **exercise** sessions, you will work through two comprehensive problems. While these **Hands-on** are designed to be challenging, we expect approximately *50% completion during class time*. The remaining work serves as take-home practice - there are no separate homework, or assignments.

You will have *two weeks* from the exercise session to **submit solutions**.

4 Assessment and Grading

4.1 Exercise Submissions (35% of Final Grade)

Exercise submissions are **evaluated** based on few specified *deliverables* described in the notebook. You will see that these are always data and analysis visualizations, in other words, **scientific plots and figures!**. After all, you are playing the role of a data scientist. It is important to know **only the requested plots will be evaluated**, not the code.

Please submit your solutions as a **PDF** file, with the plots one after the other specifying the associated task. In case one or more visualizations were not completed just write "Not completed".

Grading for each plot ranges from **3** (missing/failed submission) to **6** (perfect completion). We will use integers for each visualization: incorrect or not sufficient (3) partial completion (4) and minor imperfections (5). The grade for each exercise session is the average of the scores of the single plots, rounded to the nearest 0.25.

To ensure timely feedback, **late submissions cannot be accepted** - these automatically count as grade of 3. You will receive **grades** and **sample solutions** approximately *one week after submission*. Clarifications can be provided through Moodle Because of the *flipped* nature of the exercises, we hope we can focus Q&A during the excercises.

4.2 Final Examination (65% of Final Grade)

The course concludes with a comprehensive **Multiple Choice Question examination**. This exam is thought to be complementary to the practical skills tested in the excercises submissions and it tests your understanding of **both** theoretical concepts covered in **lectures** and emerged from the **exercises**. The MCQ format allows coverage of a broad range of topics while maintaining objective assessment criteria.

Each lecture will include **mock questions** similar to the final exam questions. These are designed to help you prepare for the final exam format and assess your understanding of the material. We will not have a classical midterm but the exercises and the will serve as a continuous evaluation and the mock questions will help you to prepare for the final exam.

The last exercise section we will also organize a **mock exam** to help you to prepare for the final exam. It will be roughly half the length of the real exam so to allow to use the rest of the time to do self-assessment and ask questions.

This is a **closed-book** exam: No lecture notes, formula sheets, or textbooks are permitted.