ECE 471 Fall 2022 - Mini Project 2

Unsupervised Analysis of Stool Microbiome in Hepatic Encephalopathy

Disclaimer

As was the case in MP1, the data and the analytical pipeline (which was based on a real experiment), have been sufficiently modified for you to learn and practice important concepts in data science. As such, the results of the analysis you perform do not accurately represent the reality of biological research. However, if you wish to learn more, you can read the representative papers provided in the References section.

Broad Problem Statement

Liver cirrhosis is a condition where a patient's liver is damaged. This damage is permanent and if severe enough can make a transplant the only treatment option [1]. Along with damage to the liver, the microbial composition of patients' gut is altered, and some patients have an accumulation of toxins in the brain leading to inflammation in the brain i.e., hepatic encephalopathy (HE) [2]. Researchers are trying to determine the relationship between the gut microbiome i.e., mircrobial composition of the gut and HE. In this project you will investigate the relationship between the gut microbiome and HE.

You'll be exploring various new data science methods in this MP and are free to use Python packages unless otherwise specified (see "Python libraries and versions" section for help). Notably, you may not use a Python package for Bayesian Networks. If you do use any additional libraries, please identify them in a markdown cell at the top of your Jupyter notebook so that we can still run your code.

Concepts you will learn and apply

- Data pre-processing and visualization
- Bayesian Networks
- Statistical analysis (Kolmogorov–Smirnov test, Multiple testing, Q-Q plot)
- Dimensionality reduction (Principal Component Analysis)
- Clustering (K-Means, Gaussian Mixture Model clustering)
- Classification

Biology Background

The gut-brain axis refers to the communication and influence between the brain and microbes in the gut [3]. This is particularly interesting because acquiring data pertaining to the gut is less difficult than gathering information on the makeup and function of the brain. Although there is strong evidence of such a connection between the brain and the gut, the exact influences they have on each other is not fully known [3].

We will be studying this connection in the context of Hepatic Encephalopathy, which is a condition in the brain that results from liver cirrhosis. A recent investigation of the gut-brain axis identified key microbes in the gut connected to Schizophrenia in the brain [4]. In that study, it was found that the abundance of different groups of microbes may change in varying ways; some may increase while others decrease, and many do not change at all.

Raw data on the microbes in the gut are the genetic sequences present in the sample – however, for our analysis, we will assume that the sequences have been analyzed in order to identify the individual microbes. The data is provided in the format of abundances. For a given sample, the abundance of each microbe is the amount of that microbe in that sample. Microbe abundances obtained from stool are reflective of the microbiomal composition of the gut.

In this project, we will study the abundance data for microbes in stool samples collected from two sets of patients: those with cirrhosis but not HE (controls), and those with cirrhosis who have developed HE (cases). **We want to identify microbes with significantly altered abundance levels between both populations and the ways in which these abundance levels change.** This information can help scientists and doctors to better direct their experimentation to understand the role of gut microbiome in the progression of HE in cirrhosis patients, and can be useful in the development of therapeutics.

Data

You have been given data for 200 patients without hepatic encephalopathy (referred to as controls) and 200 patients with hepatic encephalopathy (referred to as cases). Stool samples from these patients were processed to determine the composition of their gut microbiome, resulting in a summary of the abundance of microbes in their gut. 120 unique microbes were accounted for as a part of this analysis and the data reflects the **abundance** of each microbe for that patient. This means that the value for one microbe in one patient tells us the amount of that microbe identified from the sample.

Files

MicrobeAbundance.csv: Abundance matrix for the samples.

	<1st Microbe ID>	…	<Mth Microbe ID>	Group
<1st Patient ID>	54*	…	0.0029	case*
…	…	…	…	…
<Nth Patient ID>	47	…	0.0039	control

*The abundance of the 1st Microbe in the 1st Patient's sample is 54. The 1st Patient has hepatic encephalopathy hence belongs to the case population.

QualityControlTraining.csv: Sample collection conditions and corresponding quality scores collected from previous stool sample studies.

Collection_Method	Wait_Time	Storage_Temperature	Quality_Score
Nurse*	9.45*	-8.81*	20.85*
…	…	…	…

*A sample collected by a nurse, with 9.45 seconds wait time before storage, stored at temperature of -8.81° Celsius results in a quality score of 20.85.

QualityControl.csv: Collection conditions for samples in **MicrobeAbundance.csv.**

	Collection_Method	Wait_Time	Storage_Temperature
<1st Patient ID>	Nurse*	8.84*	-10.59*
…	…	…	…

*Sample from the 1st Patient was collected by a nurse, with 8.84 seconds wait time before storage, stored at temperature of -10.59° Celsius.

DrugResponseTrainingData.csv: Additional information for patients with HE (cases) in **MicrobeAbundance.csv.**

	effective	Sex	Age	<1st Gene ID>_before	…	<1st Gene ID>_after	…
<1st Patient ID>	1*	1*	40.52*	6.93*	…	6.79*	…

…	…	…	…	…	…	…	…

The experimental drug was effective for the 1st Patient who is a 40.52-year-old male. His expression level of the 1st Gene was 6.93 before taking the drug and was 6.79 after taking the experimental drug.

DrugResponseTestingData.csv: Information for HE patients that we want to predict the effect of the experimental drug

	<1st Microbe ID>	…	Sex	Age	<1st Gene ID>_before	…
<1st Test Patient ID>	60*	…	0*	50.70*	8.88*	…
…	…	…	…	…	…	…

*The 1st Patient (in the testing population) is a 50.70-year-old female. The abundance of the 1st Microbe in her sample is 60. Her expression level of the 1st Gene was 8.88 before taking the experimental drug.

Useful Python libraries

- numpy
- pandas
- matplotlib.pyplot
- seaborn.heatmap
- scipy.stats.norm
- scipy.stats.ks_2samp
- sklearn.decomposition.PCA
- sklearn.cluster.Kmeans
- sklearn.mixture.GaussianMixture

References

1. https://www.niddk.nih.gov/health-information/liver-disease/cirrhosis
2. Ferenci P. Hepatic encephalopathy. *Gastroenterol Rep (Oxf)*. 2017;5(2):138–147. doi:10.1093/gastro/gox013
3. Martin CR, Osadchiy V, Kalani A, Mayer EA. The Brain-Gut-Microbiome Axis. *Cell Mol Gastroenterol Hepatol*. 2018;6(2):133–148. Published 2018 Apr 12. doi:10.1016/j.jcmgh.2018.04.003

4. Zheng P, Zeng B, Liu M, Chen J, Pan J, Han Y, et al. The gut microbiome from patients with schizophrenia modulates the glutamate-glutamine-GABA cycle and schizophrenia-relevant behaviors in mice. Sci Adv 2019;5(2):eaau8317.

Task 1: Data Cleaning and Visual Inspection (30 points)

1. Bayesian Network for Quality Control (27 points)

1. **(1 point)** The Quality Score of a sample depends on the Collection Method, the Wait Time, and the Storage Temperature. The samples can be collected by either the nurse or by the patient themselves, i.e. Collection Method has values 'Patient' or 'Nurse'. The Collection Method will influence the Wait Time, as the nurses are usually more familiar with the collection process and thus more efficient. Given these information, answer the questions below:
    a. Draw a Bayesian Network that describes the relationship above.
    b. Give the factorization of the joint probability distribution represented by the Bayesian Network.

Answer the following questions using **QualityControlTraining.csv**:

2. **(2 points)** Plot the histogram of Wait Time and describe its distribution. Next, plot the histogram of log Wait Time. What might be an appropriate distribution that can fit it?
3. **(2 points)** Repeat Task 1.2, but conditioning on Collection Method this time, i.e., plot the histogram of the log Wait Time for samples collected by nurses and by patients separately. What are the mean and standard deviation of log of Wait Time when the sample is collected by a nurse?
4. **(12 points)** Assume the Quality Score of each sample is drawn from a sample-specific Gaussian distribution. The mean of the distribution is dependent on Collection Method, Wait Time, and Storage Temperature. Given a specific Collection Method (nurse or patient), the mean of the Quality Score distribution is a linear combination of (1) log of Wait Time and (2) Storage Temperature. The variance of the distribution is dependent on the Collection Method. The 1) coefficients and the intercept of the linear combination and 2) variance of the Gaussian are fixed constants for a given Collection Method. For samples collected by patients themselves, find the coefficients and the intercept of this linear combination, and also the value of the variance. Explain your estimation procedure in detail. (Hint: Use MLE)
    a. Based on the assumption above, formulate the distribution of Quality Score (Q) given Collection Method (C), Wait Time (W) and Storage Temperature (S). You can use subscript P to represent samples collected by patients and subscript N to represent nurses. You answer should be in the format of $N(\mu, \sigma)$, where μ is in terms of W and S. Use any characters to represent the unknown parameters.

b. Let's use θ_P to represent all the unknown parameters of samples collected by patients. Assume the samples were collected independently. By the MLE rule, the optimum value of θ_P is $\theta_P^* = argmax \prod_{i \in I_P} P(Q_i, C_i, W_i, S_i | \theta_P)$, where I_P are the samples collected by patients. Factorize the distribution $P(Q_i, C_i, W_i, S_i | \theta_P)$ and simplified the equation based on the Bayesian Network structure using local semantics.

c. Can you further simplify the equation for computing θ_P^* by eliminating terms irrelevant of θ_P? After that, plug in the pdf of Q to replace $P(Q_i | C_i, W_i, S_i, \theta_P)$, and further simplify the equation. What is the equation equivalent to? Can you estimate the value of θ_P^* (i.e., coefficients & intercept of the linear equation, variance of the Gaussian distribution)?

d. For samples collected by the patients: (i) Compute the mean of log Wait Time and the mean of Storage Time. (ii) Find the data that fall in a small region ($\pm$ 0.05) around the means (i.e., assuming that mean of log Wait Time is μ_W, and the mean of Storage Temperature is μ_S, find the rows satisfying $\mu_W - 0.05 \leq$ log Wait Time $\leq \mu_W + 0.05$ and $\mu_S - 0.05 \leq$ Storage Temperature $\leq \mu_S + 0.05$). What is the variance of Quality Score of those samples? (iii) How does the variance compare with the estimated variance in (c)?

5. **(5 points)** Assume in the training data, the Collection Method information is lost due to system glitches, but Wait Time, Storage Temperature, and Quality Score information is still available. Infer the most likely Collection Method for each of the entries in the training data. List your inference steps in detail (with numbers) for the first sample in the training data (first row). Display the entries (rows) where your inferred Collection Method differs from the actual data. Assume you had access to the complete training data (including Collection Method) before the glitches to estimate the parameters needed for your inference.
6. **(5 points)** Now, from **QualityControl.csv**, find out the poor quality samples. Poor quality samples have a negative Quality Score with probability greater than 5%. Use the parameters you estimated in Task 1.4. Explain your decision procedure in detail (with numbers) for the first sample in the data (first row). List the IDs of the poor-quality samples. **Remove these samples from your analysis for the remainder of this MP.**

2. Visual Inspection (3 points)

A heatmap is a visual representation where individual values contained in a matrix are represented as colors. Plot heatmaps of the abundance matrices. You're expected to plot two heatmaps - one for controls, and one for cases. The heatmaps should have samples as rows and microbes as columns. Briefly summarize your observations. Which aspects of the data are the heatmaps good at highlighting? What types of things are heatmaps less

suitable for? (Hint: Make use of the *heatmap* API in the *seaborn* package; save your plot to a local file because plotting in Jupyter Notebook is sometimes inaccurate.)

Task 2: Statistical Analysis (15 points)

Recall that the biologists wish to identify microbes with significantly altered abundance levels related to HE. A microbe's abundance is declared **altered** if the difference observed in its abundance level between controls and cases is statistically significant.

1. Kolmogorov–Smirnov (KS) Test (5 points)

1. **(1 point)** Is the KS test a parametric test or a non-parametric test? When does one want to use non-parametric tests?
2. **(2 point)** For each microbe, find the p-value of a two-sample KS test on its expression across controls vs. cases. (Hint: Make use of the *stats.ks_2samp* API in the *scipy* package.)
3. **(1 point)** What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.
4. **(1 point)** Count the number of microbes with significantly altered expression at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Summarize your answers in a table.

2. Multiple Testing (10 points)

1. **(1 point)** P-value of 0.05 is generally considered a good threshold for significant discovery. What does a p-value of 0.05 represent in our context?
2. **(1 point)** Based on the definition of p-value, if the null hypothesis is true, what distribution will the p-values follow? (Hint: Google the definition of p-value.)
3. **(1 point)** If no microbe's abundance was altered, how many significant p-values does one expect to see at alpha=0.1, 0.05, 0.01, 0.005 and 0.001 level? Compare your answers with your results in Task 2.1.3. Show the comparison in a table.
4. **(4 points)** A Q-Q (quantile-quantile) plot is used to compare two probability distributions by plotting their quantiles against each other. Say you've performed N KS tests in Task 2.1.a. Following the procedure below, plot a Q-Q plot to compare the distribution of p-values of your statistical tests (Task 2.1.1., referred to as observed p-values) with the distribution of p-values when the null hypothesis is true (Task 2.2.2.):
    1. Sample N p-values from the expected distribution in Task 2.2.2 (referred to as expected p-values).
    2. Take the -log10() of observed p-values and expected p-values.
    3. Rank observed p-values and expected p-values in ascending order separately.
    4. Take the pair of smallest p-values (one from observed p-values, one from expected p-values) and plot a point on an x-y plot with the observed p-value on the Y-axis and the expected p-value on the X-axis.

5. Repeat (iv) for the next smallest pair, for the next smallest, and so on until you have plotted all N pairs in order.
6. Add the x=y line to your plot.

5. Answer the following questions:
    1. **(1 point)** How does taking the -log10() of the p-values help you visualize the p-value distribution?
    2. **(2 points)** What can you conclude from the Q-Q plot? (Hint: Think about what it means if the Q-Q plot approximately aligns or doesn't align with the x=y line and what it implies about the null hypothesis.)

Task 3: Dimensionality Reduction and Clustering (25 points)

In this task, you will apply clustering techniques to identify subpopulations of samples.

The results in Task 2.2.e. is related to the heterogeneity of the microbiomes. For example, the cases might comprise multiple subpopulations and the relation between HE and the gut microbiome alteration might differ between or even within such subpopulation. Consequently, HE might only be related to the abundance of crucial microbes in a subpopulation of samples instead of all of them. **Thus, it is essential to first identify such subpopulations before running statistical tests.**

Identifying subpopulations based on samples' microbe abundance profiles is essentially an unsupervised clustering problem. The goal is to identify clusters of samples based on the similarities of their microbe abundance profiles. This is a difficult problem because: 1) the number of clusters is not known a priori, 2) there is usually a high level of noise in the data (both technical and biological), and 3) the number of dimensions (i.e. microbes) is large.

When working with high-dimensional datasets such as your abundance matrices, it can often be beneficial to apply some sort of dimensionality reduction method. Projecting the data onto a lower-dimensional subspace could substantially reduce the amount of noise. An additional benefit is that it is typically much easier to visualize the data in a 2 or 3-dimensional subspace, allowing us to use visual inspection as we continue analyzing the data.We will be performing Principal Component Analysis to reduce the dimensionality of the microbe abundance data.

1. Principal Component Analysis (PCA) (6 points)

The easiest way to visualize the data is by transforming it using PCA and then visualizing the first two principal components. **Note:** PCA, plotting, and calculation should be done separately for controls and cases.

1. **(2 points)** Treating microbes as features (dimensions), perform PCA on the abundance matrix. (Hint: make use of the *decomposition.PCA* API in the *sklearn* package. Select "full" for *svd_solver*.)

2. **(2 points)** Order the principal components by decreasing contribution to total variance. Plot a scree plot to show the fraction of total variance in the data as explained by each principal component. How many principal components are needed in order to explain 80% of the total variance?
3. **(2 points)** Plot a scatter plot of the abundance with only the first two components. Briefly summarize your observations.

2. Clustering (12 points)

We now perform clustering to identify the subpopulations of samples. **Controls and cases should be clustered separately.**

For all clustering algorithms below, you will need to decide (1) whether to use the original microbiome data or the transformed data from PCA for clustering, (2) whether to use all the data or some dimensions from it (the curse of dimensionality!) (3) the optimal number of clusters. **Provide reasoning for your decisions. Provide numbers, tables and/or graphs (2D or 3D) to support your reasoning.** Visualize your results by plotting a 2D scatter plot just like you did in Task 3.1.c., but with each point colored by the clusters they belong to (use different colors for different clusters).

1. **(4 points)** K-Means Clustering
2. **(4 points)** Gaussian Mixture Model Clustering
3. **(2 point)** Compare your results for different clustering methods and interpret them. Select one method for analyses that involve subpopulations for the remainder of this MP. Pay close attention to the generated clusters when choosing which results to use.
4. **(1 point)** In context, what do the clusters you have found represent?
5. **(1 point)** Based on your process for deciding the number of clusters to partition the data into, what situations or factors might result in your decision being inaccurate?

3. Identify microbes with altered abundance levels (7 points)

In this subtask, you are going to identify the set of microbes with significantly altered abundance.

For each subpopulation (cluster) in the cases, there are two possibilities:
1) Although the patients have developed HE, the makeup of their gut microbiome are not significantly different from that of patients who have not developed HE.

2) The abundance of some microbes differs significantly when compared to the gut microbiome composition of patients without HE. Thus, they have similar abundance profiles as one of the subpopulations in the controls, but not entirely the same. Consequently, they will be more similar to one of the control subpopulations compared to the other subpopulations, but not identical to any of them.

Answer the following questions (**extra credits for being creative**):

1. **(3 points)** For each case subpopulation, determine whether it has a significantly different microbiome than the controls. To get credit you must explain your method before performing the analysis. Show and explain your decision process in detail. Provide numbers, tables and/or graphs where necessary to justify your reasoning or results
2. **(2 points)** For each case subpopulation with a significantly different microbiome, identify the control subpopulation that is most similar to. Show and explain your decision process in detail. Provide numbers, tables and/or graphs where necessary.
3. **(2 points)** Identify microbes with significantly altered abundance by comparing each altered case subpopulation with its corresponding control subpopulation. Use KS test with alpha level=0.0004. This alpha level was chosen to account for potential caveats of multiple testing.

Task 4: Multi-omics drug response prediction (30 points)

Suppose there is an experimental drug to treat HE. Since the drug is not effective in every patient with HE and it has side effects, we want to be able to predict in advance which patients might benefit from the drug. Doctors know that the composition of the gut microbiome, and in particular the subtype of the microbiome profile to which the patient belongs, affects the efficacy of the drug. And drug efficacy is mediated by the expression levels of genes in the patients' serum. In addition, it is known that sex and age are also associated with the efficacy of this drug.

However, not all genes are important for the drug to work. To better understand the situation, doctors conducted a clinical trial in which they measured gene expression levels in HE patients before they took the drug and again in the same patients one month after they took the drug. (These were the same patients who participated in our microbiome study, i.e., the ***cases***.) Doctors and biologists suspect that the genes whose expression levels alter significantly after taking the drug might be those that interact with the drug and the disease. Thus, their initial expression level may be predictive of the efficacy of the drug.

In **DrugResponseTrainingData.csv**, sex, age, gene expression level before and after taking the drug, as well as whether the drug was effective were recorded. Please use this information, along with the microbiome data that you have analyzed, to construct a prediction model that predicts whether the drug will be effective. Note that the prediction model cannot directly use the gene expression levels of the patients after taking the drug, as these data will not be available for the new patients for whom we wish to make a prediction.

Apply your most confident model to a new set of HE patients. Their microbiome composition, sex, age, and gene expression levels (before taking the drug) are recorded in **DrugResponseTestingData.csv**. Save your prediction in a .csv file named **DrugResponseTestingData_prediction.csv** with two columns and no header. Patient_ID should be the first column and your binary prediction for each of the patients be the second.

Explain in your Jupyter Notebook and in your slide presentation on how you integrate the insights you got from the doctors and biologists. For example, do you want to perform dimensionality reduction? How do you perform feature selection (if it's necessary)? What algorithms have you tried and how well each of the performed and why? Also explain what methods or procedure you use to arrive at your **most confident** model.

Deadlines

Checkpoint #	Deadline	Tasks	Requirements
1	10/12 11:59:59 PM	Tasks 1, 2	· Canvac submission · .ipynb with tasks 1 and 2 completed · Powerpoint presentation for tasks 1 and 2 (.pdf)
1.5	10/24 11:59:59 PM	-	· Fill out progress report via Google form
2 (Final)	10/31 11:59:59 PM	Tasks 1-4	· Canvas submission · .ipynb with all tasks completed · Powerpoint presentation for all tasks (.pdf)

Submission Requirements

Please provide a single .ipynb file. Please label each section, task, and subsection accordingly.

- Write your names and NetIDs of group members in the beginning
- Explain all your work (include the code with comments)
- Write down the equations that are being used (for partial credit)
- All the charts should be appropriately formatted by showing the legend, axes labels, and chart title
- Each question answered should include the code you used to achieve the needed charts and/or tables and an explanation/interpretation

Please also prepare a powerpoint with your results for checkpoints 1 and 2. Templates for these powerpoints will be released closer to the due date for each checkpoint.

All submissions aside from the progress reports are done on Canvas. Please don’t zip your files - upload them as separate files. Late submission policy is applicable. One submission per group.

Academic Integrity

When completing this project, please ensure that your submissions reflect your own work and ideas. Academic integrity is taken seriously, and any issues will be dealt with strictly to ensure a fair course for all students. All submissions will be tested at the end of the MP to check for plagiarism of both answers and code.