ECE/CS 498 DS Spring 2020 - Mini Project 2
Unsupervised Stool Sample Analysis in Hepatic Encephalopathy

Disclaimer

As was the case of MP1, the data and the analytical pipeline (which was based on a real experiment), has 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, representative papers are 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, some cirrhosis patients have an accumulation of toxins in the brain known as hepatic encephalopathy (referred to as HE) [2]. Researchers are trying to determine the relationship between the condition of the digestive system and HE. In this project you will investigate the relationship between the microbial composition of the digestive tract 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, t-SNE)
- Clustering (K-Means, Gaussian Mixture Model clustering, Hierarchical clustering)

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. At this stage, the data is usually provided in the format of relative abundance. For a given sample, the relative abundance of each microbe is the proportion of the amount of that microbe to the total amount of all microbes present. Since stool is a product of the digestive tract, we can analyze stool samples to understand the microbes present in the patient's gut.

In this project, we will study relative abundance data for microbes in stool samples collected from two sets of patients: those with cirrhosis but not HE (HE0 population), and those with cirrhosis who have developed HE (HE1 populations). **We want to identify microbes with significantly altered abundance levels between both populations and the ways in which these abundance levels change.** Although such results will not directly answer the question of how HE develops in cirrhosis patients, they can help scientists and doctors to better direct their experimentation on how to map out the gut-brain axis.

Data

You have been given data for approximately 750 patients without hepatic encephalopathy (referred to as HE0 samples) and approximately 750 patients with hepatic encephalopathy (referred to as HE1 samples). Stool samples from these patients were processed to determine the composition of their gut microbiome, resulting in a summary of the microbes in their gut. About 150 unique microbes were accounted for as a part of this analysis and the data reflects the **relative abundance** of each microbe for that patient. This means that the value for one microbe in one patient tells us that the proportion of that microbe in all microbes identified from the sample.

Files

MP2_template.ipynb: Provided template for MP2 notebook.

RelativeAbundance_HE0.csv: Relative abundance matrix for HE0 samples.

Microbe Name	HE0Sample_0	…	HE0Sample_N
<Microbe 1 Name>	0.0042*	…	0.0029
…	…	…	…
<Microbe N Name>	0.0013	…	0.0039

*relative abundance of microbe 1 in sample 0 is 0.0042.

RelativeAbundance_HE1.csv: Relative abundance matrix for HE1 samples, same format as **RelativeAbundance_HE0.csv.**

Note: The microbe name is a unique identifier. A given microbe name will only appear once in **RelativeAbundance_HE0.csv** and in **RelativeAbundance_HE1.csv.** Microbes are sorted alphabetically based on their names in both files.

QualityControl.csv: Sample collection conditions and corresponding quality records collected from previous stool sample studies.

strtmp	coll	labtime	cont	qual
cold*	nurse*	short*	low*	good*
…	…	…	…	…
…	…	…	…	…

*A sample collected by a nurse, stored at cold temperatures, and processed in a short amount of time in the lab results in a low chance of contamination and good quality data.

BayesInferenceHE0.csv: Collection conditions for samples in **RelativeAbundance_HE0.csv.**

SampleName	strtmp	coll	labtime
HE0Sample_0*	cold*	nurse*	short*
…	…	…	…
HE0Sample_N	…	…	…

*HE0Sample_1 was collected by a nurse, stored at cold temperatures, and process in a short amount of time in the lab.

BayesInferenceHE1.csv: Collection conditions for samples in **RelativeAbundance_HE1.csv**, same format as **BayesInferenceHE0.csv.**

Python libraries and versions

Please use Python 3 with the following library versions or newer for all code in this assignment.

- Numpy – 1.15.1
- pandas – 0.25.3
- matplotlib.pyplot – 2.2.3
- seaborn.heatmap – 0.9.0
- scipy.stats.ks_2samp – 1.1.0
- sklearn.decomposition.PCA – 0.20.3
- sklearn.manifold.TSNE – 0.20.3
- sklearn.cluster.Kmeans – 0.23.0
- sklearn.mixture.GaussianMixture – 0.20.3
- sklearn.cluster.AgglomerativeClustering – 0.20.3

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

0. Getting Started With the Data

Answer the following questions based on **RelativeAbundance_HE0.csv**:
1. In the context of statistical analysis, why do biologists need multiple samples to identify microbes with significantly altered abundance?
2. How many samples were analyzed?
3. How many microbes were identified?

1. Bayesian Network for Quality Control
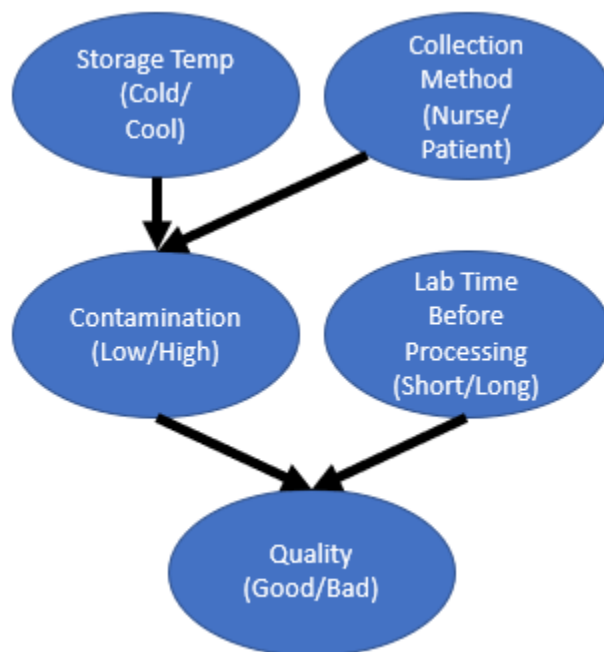
a. Give the factorization of the joint probability distribution.
b. Count the number of parameters needed to define the conditional probability distribution of the Bayesian Network for quality control.
c. Show the conditional probability tables P(Quality|Contamination, Lab Time), P(Contamination|Storage Temp, Collection Method), P(Storage Temp), P(Collection Method), P(Lab Time) for the above network. Training data is provided in **QualityControl.csv**.
d. Calculate P(Quality|Storage Temp, Collection Method, Lab Time) for all possible values of Quality, Storage Temp, Collection method, and Lab Time. Show your calculation.
e. Use the calculated conditional probabilities and the collected data **BayesInferenceHE0.csv, BayesInferenceHE1.csv** to determine the quality of the analyzed stool samples given data on the Storage Temperature, Collection Method, and

Lab Time Before Processing. Report bad quality samples. Drop bad quality data for the following analyses.

Use the good-quality data for the remainder of the MP.

2. Data Standardization

1. As we are being provided **relative** abundance data for microbes in the stool samples, the sum of all the values for each sample should be 1. Before we can begin analyzing our data however, it is important that we check this to ensure that the provided data is as expected. Please verify that we have indeed been provided with relative abundance data. If the provided data has samples which do not follow these constraints, please identify and remove those samples.

2. What are the benefits and drawbacks to using relative abundance data? Is there information that we lose when the normalization is performed?

Use the normalized data for the remainder of the MP.

Other normalization techniques include subtracting the mean, taking the z-score, etc. One should choose the normalize technique carefully based on one's understanding of the domain and the nature of the analyses.

3. Visual Inspection

A heatmap is a visual representation where individual values contained in a matrix are represented as colors. Plot heatmaps of the relative abundance matrices. You're expected to plot two heatmaps - one for HE0 samples, and one for HE1 samples. The heatmaps should have microbes as rows and samples 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

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 HE0 samples and HE1 samples is statistically significant.

1. Kolmogorov–Smirnov (KS) Test

a. For each microbe, find the p-value of a two-sample KS test on its expression across HE0 samples vs. HE1 samples. (Hint: Make use of the *stats.ks_2samp* API in the *scipy* package.)

b. What is the null hypothesis of the KS test in our context? Use one microbe as an example to explain your answer.
c. 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

a. 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?
b. 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.)
c. 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.c. Show the comparison in a table.
d. 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.a., referred to as observed p-values) with the distribution of p-values when the null hypothesis is true (Task 2.2.b.):
   i. Sample N p-values from the expected distribution in Task 2.2.b (referred to as expected p-values).
   ii. Take the -log10() of observed p-values and expected p-values.
   iii. Rank observed p-values and expected p-values in ascending order separately.
   iv. 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.
   v. Repeat (iv) for the next smallest pair, for the next smallest, and so on until you have plotted all N pairs in order.
   vi. Add the x=y line to your plot.
e. Answer the following questions:
   i. How does taking the -log10() of the p-values help you visualize the p-value distribution?
   ii. 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

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

The results in Task 2.2.e. is related to similarity between samples. For example, the ~700 HE0 samples might comprise multiple subpopulations and the relation between HE and the gut microbiome might differ between or even within such. 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. That 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 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 relative 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. Therefore, we will be performing Principal Component Analysis and t-Distributed Stochastic Neighbor Embedding to reduce the dimensionalities of the microbe abundance data.

1. Principal Component Analysis (PCA)

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 HE0 samples and HE1 samples.

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

    b. 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 30% of the total variance?

    c. Plot a scatter plot of the relative abundance with only the first two components. Briefly summarize your observations.

2. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE is a popular dimensionality reduction technique for visualization. t-SNE models each high-dimensional object by a two- or three-dimensional point such that similar projects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability. We are not expecting you to know t-SNE in detail, but it is useful to know how to run t-SNE with a package and interpret the results.

    a. Using the given samples and treating microbe abundances as features (dimensions), perform t-SNE to visualize the data in 2D. (Hint: make use of the *manifold.TSNE* API in the *scikit-learn* package with **parameter random_state=42**) **t-SNE and plotting should be done for HE0 samples and HE1 samples separately.** Repeat the same t-SNE and

plotting using the values 62 and 82 for random_state. Briefly summarize your observations.

b. Discuss the similarities and differences between results when clustering with PCA and t-SNE.

3. Clustering

We now performing clustering to identify the subpopulations of samples. For the purpose of this project, **clustering should be performed after projecting the data into 2D using PCA as t-SNE is dependent on a random initialization seed** (Task 3.1.a.).

HE0 samples and HE1 samples should be clustered separately. 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). For all clustering algorithms below, **you will need to decide the optimal number of clusters by yourself and reason about it.** Provide numbers, tables and/or graphs to support your reasoning.

a. K-Means Clustering
b. Gaussian Mixture Model Clustering (Hint: You might run into numerical issues if you use the GMM implementation in *scikit-learn*. You might want to try scaling up the data 100 or 1000 times if your results look weird.)
c. Single and Complete Linkage Hierarchical Clustering
d. Discuss the differences between the single and complete linkage hierarchical clustering methods. Do you see any major differences in the generated clusters? Is there anything about our data which affects if we see a difference between the two linkage options?
e. Compare your results for different clustering methods and interpret them. Select the results from one of the clusters for the following analyses. Pay close attention to the generated clusters when choosing which results to use.
f. In context, what do the clusters you have found represent? What are some factors which could account for this type of clustering pattern?
g. 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?

Task 4: Interpreting Your Results

In this task, you are going to identify the set of microbes with significantly altered abundance and the way in which they change.

1. Identify microbes with altered abundance levels

For each subpopulation (cluster) in the HE1 samples, there are two possibilities:
1) Although the patient has HE, the makeup of their gut microbiome is not significantly different from those of patients who have not developed HE.

2) The relative abundance of some microbes differs significantly when compared to the gut microbiome composition of patients without HE. Thus, they have similar relative abundance profiles as one of the subpopulations in the HE0 samples, but not entirely the same. Consequently, they will be more similar to one of the HE0 subpopulations compared to the other subpopulations, but not identical to any of them.

Based on the above information, answer the following questions:
  a. For each HE1 subpopulation, determine whether it has a significantly different microbiome than the HE0 samples. 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.
  b. For each HE1 subpopulation with a significantly different microbiome, identify the HE0 subpopulation that is most similar to. Show and explain your decision process in detail. Provide numbers, tables and/or graphs where necessary. When performing this step keep in mind that 5-20 microbes should have a significantly altered abundance.
  c. Identify microbes with significantly altered abundance by comparing each altered HE1 subpopulation with its corresponding HE0 subpopulation. Use KS test with alpha level=0.0000025. This alpha level was chosen to account for multiple testing caveats implied in Task 2.

2. Identifying how abundance changes for each microbe in our data

Although we may have identified the microbes which change, our analysis is not complete. We will now further analyze these microbes to see how they change and see if we can identify any further patterns:

  a. Which of the microbes that you identified show an increase of relative abundance in the HE1 sample? Do any show a decrease?
  b. Are there any taxonomical relationships between the microbes with altered abundance? If so, identify these groupings.
     Hint: The long microbe names in the data include information on the taxonomic groupings for each microbe. Try searching to see which names correspond to what taxonomy levels.

Deadlines

Checkpoint #	Deadline	Tasks	Requirements
0.5	3/2 11:59:59 PM	-	• Fill out progress report via Google form
1	3/13 11:59:59 PM	Tasks 1, 2	• Compass2g submission • .ipynb with tasks 1 and 2 completed • Powerpoint presentation for tasks 1 and 2 (.pdf)
1.5	3/25 11:59:59 PM	-	• Fill out progress report via Google form
2 (Final)	3/30 11:59:59 PM	Tasks 1-4	• Compass2g 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 Compass2G. 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.