CSE 5520 Fall 2021
Homework 4 (Due 11:59 pm, Sunday, October 3, 2021 at HuskyCT)

Visualization of t-Test, Clustering, Heatmap and Bayesian Inference

This homework is to help you practice with some data visualization fundamentals. You are expected to use these visualization techniques and others in your final project. You are required to do this exercise in Python. All plots/graphs must have titles and x-y coordinate tick labels.

Part 1: t-Test

Consider again the gene expression data sets introduced in Homework 3. Your analysis is limited to comparing the Dementia group (GSE44768_CR_alz_female_reduced.csv) and the Non-Dementia group (GSE44768_CR_nd_female_reduced.csv). Your study goal is to construct and contrast a 2 x 2 plot capturing various t-test outcomes for 2 genes, LAMP2 and BSN. In terms of the columns of the 2 x 2 plot, the first column is to show two histograms in one plot (red for Dementia and blue for non-Dementia for one gene) similar to what you practiced in Homework 3. The second column is to show t-Test PDF. The steps given below is to guide you how to do this exercise.

Step 1. Create and show two histograms in one plot for the gene LAMP2 with different colors, red for the Dementia group and blue for the Non-Dementia group into one plot. It should be the first subplot of the first 1 x 2 plot.

Step 2. Perform t-Test (Hint, unequal variance, unequal sample size) and show PDF. In the PDF plot, indicate the critical region (significance level $\alpha = 0.05$) by placing two dots, arrows or lines near the x-axes. You can even color the critical region (area) on the plot if you know how to do this **for extra credits** (Note visual communication is the theme of this course). Coloring the critical region is not required. Then display t_{obs} by placing a different colored dot, arrow or line near the x-axes. Using a markdown cell, state if you would reject H_0 or not. H_0 is "The mean values of gene expression between the Dementia group and Non-Dementia group among female patients are **not** different" and H_1 is "The mean values of gene expression between the Dementia group and Non-Dementia group among female patients are different (one could be higher or lower than the other)". This t-Test's PDF is the second subplot of the first 1 x 2 plot. The plot title for the PDF should include the calculated DF (degrees of freedom).

Step 3. Repeat Steps 1 and 2 for the gene BSN. This is to generate the second 1 x 2 plot.

Step 4. Lastly, stack two 1 x 2 plots you created in Steps 1, 2 and 3 "top to bottom" to create one 2 x 2 plot and display it. Use a markdown cell and explain how to interpret/contrast the plots you generated. For example, in which case you would reject/accept H_0 and why.

Part II: Clustering and Heatmap

Consider the pre-processed Alzheimer data set GSE44768_CR_40_54_Combined.csv which is available at HuskyCT's Data folder. MS Excel version is also available that differentiates samples by color highlights.

Step 1. Import and show a 40 x 54 gene expression value matrix. The row ids (sample name) should make it clear how this matrix was created. Only 10 samples from different patients are included in this file. Here alz/nd means dementia or no and f/m denotes gender.

Step 2. Perform a sample-wise hierarchal clustering using Ward method for the linkage algorithm. Note that using Ward method can be easily done by choosing the right option from the library. The outcome should exhibit how the 40 samples from four subgroups (Male Dementia/ Male Non-Dementia / Female Dementia / Female Non-Dementia) are clustered. Using a markdown cell, discuss how you would interpret the outcome, i.e., is it what you expected to see or not, i.e., samples from each of the four groups are clustered?

Step 3. Perform a gene-wise hierarchal clustering using Ward method for the linkage algorithm. It should exhibit how the genes are clustered independent of sample group membership. Using a markdown cell, discuss how you would interpret the outcome, i.e., do you notice formation of any subgroups of genes behaving similarly across samples?

Step 4. Perform the ultimate hierarchal clustering by performing both column-wise and gene-wise clustering. Again you use Ward method for the linkage algorithm. Using a markdown cell, discuss how you would interpret the outcome. For example, can you isolate a group of meaningful genes cutting across gender (M/F) and/or disease phenotype (Dementia/Non-Dementia)?

Part III: Bayesian Inference

Consider two people, Joe and Jim, who went to cancer screening at different clinics, Joe for prostate cancer screening and Jim for breast cancer screening. Male breast cancer is rare ("less than 1% of all breast cancers occur in men", https://www.breastcancer.org/symptoms/types/male_bc), but prostate cancer is common ("About 6 cases in 10 are diagnosed in men who are 65 or older", https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html). Both heard that their cancer screenings turned out positive. In this exercise, assume P (Cancer=Male_breast) = 0.01 and P (Cancer=Male_prostate) = 0.6. Also assume TPR and TNR for prostate cancer screening are 0.75 and 0.70 and TPR and TNR for breast cancer screening are 0.85 and 0.80, respectively.

Step 1. Create a contingency table for Bayesian inferencing for each individual. You should show two tables, one for Joe and one for Jim.

Step 2. This time, convert each contingency table you created in Step 1 into a grid-based plot designed to visualize Bayesian inference. Show two 20x20 grid-based illustrations "side by side" to contrast the two cases. You are illustrating Bayesian inferencing using the population size of 400 people in each case, following the discussion we had in class. Coloring of dots properly should differentiate TP/FP/FN/TN cases (i.e., treating dots as people belong to each of these four cases). Suggested colors for dots are: TP – red, FP – blue, FN – maroon, and TN – gray following the convention used in "Example 2 – Drug testing" in lecture slide. Boxing the dots using the code given in the class to label the cancer population (TP $\cup$ FN) and/or non-cancer population (TN $\cup$ FP) can be done **for extra credit.** Using the boxing to label the groups is not required. Make sure the titles of plots include person name and P (Cancer=yes | test=positive) , i.e., the probability that the person has cancer upon hearing the positive screening outcome news.

Step 3. Using a markdown cell, discuss what you can say with the two 20x20 grid-based plots you are contrasting, e.g., the impact of prior vs. likelihood in computing the posterior.

Various Useful Resources

Computing Degree of Freedom (DF):

https://stackoverflow.com/questions/49473757/python-degrees-of-freedom
There is an error in this URL post. The corrected and "augmented" version is here. Setting the variables A and B is included:

```
from statistics import mean
import math

def stdev(X):
    m = mean(X)
    return math.sqrt(sum((x-m)**2 for x in X) / len(X)-1)

def degreesOfFreedom(X, Y):
    s1 = stdev(X)   # standard deviation of sample 1
    s2 = stdev(Y)   # standard deviation of sample 2
    s1_sq = (s1**2) # variance of sample 1
    s2_sq = (s2**2) # variance of sample 2
    n1 = len(X) # length of sample 1
    n2 = len(Y) # length of sample 2
    df = ((s1_sq/n1)+(s2_sq / n2))**2 / ((s1_sq/n1)**2 / (n1 - 1) + (s2_sq/n2)**2 / (n2 - 1))
    return(df)

A = [1, 2, 3, 4, 5, 6]    # includes values for Alzheimers'
B = [5, 6, 7, 8, 9, 10]   # includes values for Non-Dementia

print('Degrees of freedom for Student-t distribution: ' + str(degreesOfFreedom(A, B)))
```

Plotting t-test PDF:

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html

You can use 0.001 and 0.999 instead of 0.01 and 0.99 to stretch tails.

```
x = np.linspace(t.ppf(0.001, df),
                t.ppf(0.999, df), 100)
```

Calculating Critical Region:
https://machinelearningmastery.com/critical-values-for-statistical-hypothesis-testing/

Note: You use two tail t-test since you do not care if test sample distribution is larger or smaller than the control sample distribution; you care only different or not. In case of one tail, you set p=0.95. In case of two tail test, you set p = 0.975 for the critical region calculation.

Clustering and Heatmap:
https://seaborn.pydata.org/generated/seaborn.clustermap.html

https://www.geeksforgeeks.org/hierarchically-clustered-heatmap-in-python-with-seaborn-clustermap/

https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html#scipy.cluster.hierarchy.dendrogram

You upload your Jupyter notebook in HuskyCT. The file name should be of the following format: HWn_Doe where n is the homework number and Doe denotes the last name.

HWs and Projects, 5% penalty for one day late submission. No acceptance after 5 days late. Extension is allowed only with the supporting medical record.