MSA 8040 Data Management for Analytics: Project 2

Houping Xiao, Georgia State University

October 13, 2022

Requirements (Please read it carefully!):

1. ***Please keep it confidential and do not distribute it outside our cohort!***
2. ***Report must be given for each group.***
3. ***Report should be submitted on iCollege before 11:59 pm on 11/06/2022.***
4. ***This is group-based project. Only one group member need to submit the files.***
5. ***Individual-based peer evaluation form. Each group member need to submit their own evaluation for the other group members in your group.***

   Peer-Evaluation Form

   To better achieve fairness in the class, at the end of the course you will be asked to evaluate yourself and the other members of your group on completing the project. These ratings are used for gauging team members' contributions. The grade you and your group members receive will depend in part on these peer evaluations. Rate each member based on the following criteria: (1) participation in group activities, (2) quality of work, (3) quantity of work, (4) finishing assigned work on time, and (5) ability to work as a team member. Please use the following scale to assign scores:

5	Exceptional effort, above and beyond the call of duty
4	Above average effort
3	Normal effort (this is the expected score!)
2	Below average effort
1	Unacceptable effort

   Then, submit the following note to the instructor on iCollege:

   Your Name:________________ Score:____ Reasons:____________
   Team Member #2:____________ Score:____ Reasons:____________
   Team Member #3:____________ Score:____ Reasons:____________
   Team Member #4:____________ Score:____ Reasons:____________
   Team Member #5:____________ Score: ____Reasons:____________

   Note: Please include a brief reason for any group member. I expect everyone to be thoughtful and diligent in completing this evaluation. Your final project grade will be biased by the pear-evaluation scores fro your group members. For instance, you could be graded as ZERO for the project if you receive "1"s from all other group members.

1 Database

Use your database built for **Project 1** to answer the following questions.

1.1 Part I [60%]

1. For probes belonging to GO with a given id value, write a stored procedure or function, named **AvgDifference(id)**, to calculate the difference between the average of the expression values of patients with "ALL" and the average of expression values of patients without "ALL". For instance, **AvgDifference("0012502")** is to calculate the difference between the average of the expression values of patients with "ALL" and the average of expression values of patients without "ALL". [**30%**]

2. For probes belonging to GO with a given id value, write a stored procedure or function, named **GeneStatistics(id, diseaseName)**, to calculate the number of distinct expression values, average of the expression values, and the summation of the expression values among patients with a given disease. For instance, **AvgCalculation("0007154", "ALL", Count, Avg, Sum)** calculates the number, average, and summation of the expression values among patients with "ALL", and save the results into count, avg, and sum respectively.[**40%**]

1.2 Part II [30%]:

Given a specific disease and a given gene, please tell whether the gene is one informative gene for the disease. You can write a stored procedure or function in MySQL workbench, OR you can use MySQL to query and use other software like Python to calculate the t-statistic.
The steps to tell whether a gene is informative to disease is listed as follows. For example, suppose we are interested in the cancer "ALL".

1. Find all the patients with "ALL" (i.e., Group A), while the other patients serve as the control group (i.e., Group B).

2. For the gene, calculate the t-statistics for the expression values between Group A and Group B.

3. If the p-value of the t-test is smaller than 0.01, this gene is regarded as an "informative" gene.

1.3 Part III: Bonus Question [20%]:

Identify all the informative genes and use informative genes to classify a new patient. There are five test cases in test_samples.txt are given in the data.
For example, given a new patient P_N, we want to predict whether he/she has "ALL".

1. Find the informative genes with respect to "ALL"

2. Find all the patients with “ALL” (i.e., Group A)

3. For each patient P_A in Group A, calculate the correlation r_A of the expression values of the informative genes between P_N and P_A

4. Patients without “ALL” serves as the control group (i.e., Group B)

5. For each patient P_B in Group B, calculate the correlation r_B of the expression values of the informative genes between P_N and P_B

6. Apply t-test on r_A and r_B, if the p-value is smaller than 0.01, the patient is classified as “ALL”