CS 4491/CS 7990- Homework 3

In HW3, we will infer a gene regulatory network from gene expression data and make a ROC plot.

Download the gene expression data in the course web page, where there are 500 samples and each sample has 10 gene expression. There are two tasks in HW3.

Task 1: Gene regulatory networks inference based on the correlation-based approach.

- Required for both undergraduate/graduate students
- Dataset:
    - Gene_expression_1.csv: contains gene expression data for task 1
    - Adj_1.csv: contains adjacency matrix of ground truth for task 1

Task 2: Gene regulatory networks inference based on the regression-based (LASSO) approach

- Required for graduate students only
- Dataset:
    - Gene_expression_2.csv: contains gene expression data for task 2
    - Adj_2.csv: contains adjacency matrix of ground truth for task 2

Bonus:

- If an undergraduate student complete task 2 as well as task 1, bonus points (extra 50 points) will be given.
- If you implement LASSO by yourself (instead of using a library function), additional bonus points (extra 50 points) will be given.

For the correlation-based approach,

1. Load the gene expression data (Gene_expression_1.csv) and the ground truth adjacency matrix (Adj_1.csv).
2. Compute pairwise correlation matrix, and show the matrix. E.g., see Fig. 1.
3. Given the range of threshold (e.g., 0, 0.1, 0.2, 0.3, …, 0.9, 1), compare the adjacency matrices between the network and the ground truth.
4. Compute a confusion matrix for each threshold
5. Compute TPR and FPR for each threshold
6. Make a ROC plot. E.g., see Fig. 2

```
             [,1]         [,2]         [,3]         [,4]         [,5]         [,6]         [,7]         [,8]         [,9]         [,10]
 [1,] 0.0000000000 1.731708e-01 1.054191e-04 4.215201e-04 3.616903e-01 6.628425e-01 0.3238827683 6.998978e-04 1.313074e-01 4.178738e-01
 [2,] 0.1731708398 0.000000e+00 2.612879e-04 7.774737e-07 5.903489e-01 2.824813e-01 0.5282463659 1.042277e-03 6.778354e-01 6.582586e-01
 [3,] 0.0001054191 2.612879e-04 0.000000e+00 2.448600e-04 3.001712e-04 2.833031e-05 0.0001954879 3.948046e-04 2.132722e-04 1.453577e-05
 [4,] 0.0004215201 7.774737e-07 2.448600e-04 0.000000e+00 4.248804e-05 7.924904e-04 0.0002506381 6.551356e-03 1.646264e-04 1.154319e-06
 [5,] 0.3616903367 5.903489e-01 3.001712e-04 4.248804e-05 0.000000e+00 6.001508e-01 0.5220584368 1.316574e-03 4.121609e-01 8.870413e-01
 [6,] 0.6628425202 2.824813e-01 2.833031e-05 7.924904e-04 6.001508e-01 0.000000e+00 0.5010528062 2.813857e-04 1.940055e-01 6.664619e-01
 [7,] 0.3238827683 5.282464e-01 1.954879e-04 2.506381e-04 5.220584e-01 5.010528e-01 0.0000000000 2.508563e-04 3.773563e-01 5.754074e-01
 [8,] 0.0006998978 1.042277e-03 3.948046e-04 6.551356e-03 1.316574e-03 2.813857e-04 0.0002508563 0.000000e+00 3.560426e-05 1.441785e-03
 [9,] 0.1313074479 6.778354e-01 2.132722e-04 1.646264e-04 4.121609e-01 1.940055e-01 0.3773562754 3.560426e-05 0.000000e+00 4.638732e-01
[10,] 0.4178738114 6.582586e-01 1.453577e-05 1.154319e-06 8.870413e-01 6.664619e-01 0.5754074300 1.441785e-03 4.638732e-01 0.000000e+00
```

Figure 1. Correlation matrix

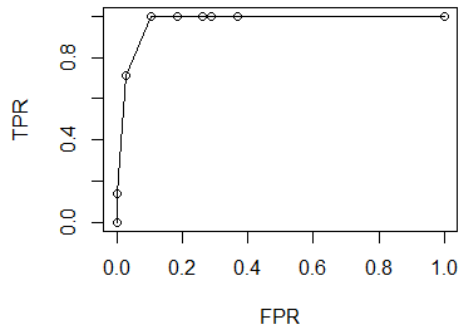

Figure 2. ROC in Task 1

For the regression-based approach,

1. Load the gene expression data (Gene_expression_2.csv) and the ground truth adjacency matrix (Adj_2.csv).
2. Given the set of lambdas, {0, 0.001, 0.005, 0.001, 0.01, 0.05, 0.1, 0.5, 1, 10, 100}, construct the adjacency matrix with the coefficient result of LASSO
3. Compare between the network and the ground truth.
4. Compute a confusion matrix for each threshold
5. Compute TPR and FPR for each threshold
6. Make a ROC plot.

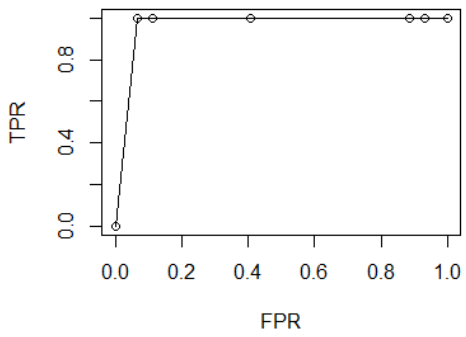

Figure 3. ROC in Task 2

Submission:

You have to submit the followings to D2L:

1. MS word file
   - Describe what you did for the homework assignment.
   - Specify which bonus tasks you completed on top if you did.
   - Include the two figures
   - Include the correlation matrix in Task 1.
   - Include the values of TPR and FPR for each task. For example, there are 11 TPR and FPR in Task 2, because we tested 11 lambdas for ROC.

2. Source code file(s)
   - Any languages, but recommend R, Python, or Matlab
   - Must be well organized (comments, indentation, …)

Deadline:

You have to submit HW3 by **Wednesday, March 15, 2017**. Late assignments will be accepted up to 24 hours after the due date for 50% credit. Assignments submitted more than 24 hours late will not be accepted for credit.