Network Inference Assignment
BCB570

1. Bagging: Bootstrap Aggregrating
a. Describe the bagging process in your own words.
b. Use R or Python to generate a sample of 100 instances of a normally distributed random variable with a mean of 2 and a variance of 8. Estimate the mean and variance of the data from the entire sample of 100.
c. Write a program that subsamples the data (with replacement) into 20 groups of 10 and estimates the mean and variance for each subsample. Plot histograms of the sample mean and variance for each subsample. Use the histograms to estimate the sample parameters.

2. Association Networks
a. Describe the basic methodology for creating association networks by describing the key steps.
b. Create a novel association algorithm using a combination of the methods that we described in class for omics data that you work with. Justify your selection of association method, thresholding, and filtering methods given your organism and data type.

3. Data Processing Inequality

3. Data Processing Inequality (DPI)

Suppose we have a probability model described as a Markov Chain:

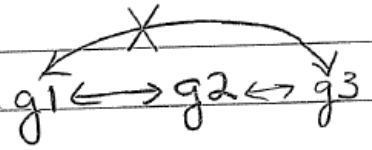
$$X \rightarrow Y \rightarrow Z$$

where $X \perp Z \mid Y$, then it must be that

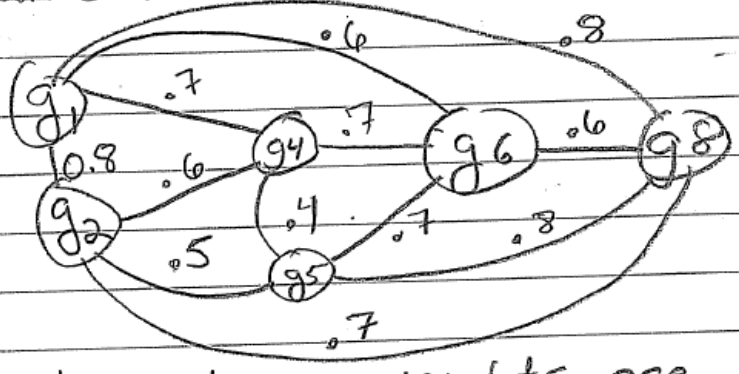
(1) $I(X, Y) \geq I(X, Z)$.

a. Prove that (1) is true using the chain rule and the conditional independence condition ($X \perp Z \mid Y \Rightarrow I(X, Z \mid Y) = 0$)

b. How is the DPI used to simplify Mutual Information Networks?

c. 

If $I(g_1, g_3) \leq \min(I(g_1, g_2); I(g_2, g_3))$, then can eliminate link between g1 and g3



Given the edge weights are $I(g_i, g_j)$ Simplify the network.

Network Inference Assignment
BCB570

4. Network Inference

Using the datasets below use two of the methods given below, or for extra credit find a new hybrid method from the recent literature (2014 or later), to learn the genetic regulatory network. Possible methods are WGCNA, TIGRESS, GENIE3, ARACNE/CLR.

a. Describe how the method finds associations between genes and how you are thresholding the network in each case.
b. Data set 1: experimental data from DREAM 3 competition with the gold standard for checking result. Compare method accuracy using Precision/Recall Curves.
c. Blind data set 2. Come up with your best estimate for the ecoli data set given what you learned in part b. Your results should be a list of undirected edges. We will assess it using the CURRENT version of RegulonDB.

5. After the main assignment has been turned in, we will create an ensemble estimate of the different blind data set results. Design an algorithm with a weighted voting scheme that uses the wisdom of crowds that we can implement.