CISC 889 Modeling and Simulation
Fall, 2017
Homework 2

Due date: October 17, 2017

In this assignment, you are asked to reconstruct gene regulatory network from gene expression data by implementing Bayesian networks and to make inference about gene expression.

Regulatory network reconstruction. In this part, you are going to implement the learning algorithm based on maximum likelihood.

For any given graph G, a score(G: D) is computed from the probability as the following

$$score(G: D) = \log P(D|G)$$

For this assignment, one simplification is to impose an upper bound k=3 for the number of parent nodes for any given node in the graph.

The pseudo code of the algorithm using greed search is given below.

Input:

G_0 // initial network structure
D // a data set

Algorithm:

$G_{best} = G_0$
repeat
    $G = G_{best}$
    Randomly select a pair of nodes
    foreach operator o (each edge addition, deletion, or reversal on G)
        G' = o(G) // G' is the new graph
        if G' is cyclic **next**
        if score(G': D) > score(G_{best}: D)
            G_{best} = G'
until G==G_{best} // no change in structure improves score

Output: G

In implementing the above algorithm, you need two subroutines. Subroutine one is to update the conditional probability table (CPT) for each node given a graph and the data, using ML procedure. Subroutine two is to test if a given graph is cyclic.

Subroutine one:

Input: D // data
G // graph

Algorithm:

foreach node v in G
obtain the parent nodes Pa(v)
calculate M[x, **u**] // number of times v = x and Pa(v) = **u**
calculate M[**u**] = $\sum_{x=0,1}$ M[x, **u**]

for x = 0 and 1
calculate CPT (v) = P(v = x| Pa(v) = **u**)
= M[x,**u**] / M[**u**]

Output: CPT(v) for every v in G.

Subroutine two: refer to any algorithm text, e.g., CLRS.

Test and evaluation. In this part you test the network by predicting if the expression level of the fifth gene g3 is up (1) or down (0), given the expression levels of other genes.

For each row in the testing data,

Prediction (g5) = 1 if P(g5 = 1 | Pa(g5)) > 0.5
= 0 otherwise

Calculate the prediction accuracy = # of correct prediction / # of total predictions.

The both training and test files contain tabular data. There are five columns corresponding to five genes. Each row gives the expression levels (0 or 1) for these five genes. The goal is to infer regulatory network for these genes. The training and testing data can be downloaded from

The train and test data can be downloaded from Sakai: hw2_train.dat, hw2_test.dat.

Repeat the training and testing five times. Report G_{best},CPT (G_{best}), and Accuracy

What to hand in?

1. Your code along with a readme file about how to compile and run it.
2. Summary of results in a PDF file.