GENOME 541, Spring 2012
Problem Set #1

(Due Apr 21th 11:59pm)

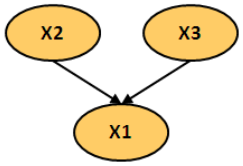
---

1. **[40 points] Parameter estimation in Bayesian networks**

   In this question, we will consider a Bayesian network representing a regulatory network among N genes. Denote by $X_1, ..., X_N$ *variables* representing the expression levels of these N genes, and by x_i a *value* assigned to X_i. For simplicity, we assume that each X_i's value is discretized to two levels, i.e., $x_i \in \{high, low\}$. Let's assume that the structure of the Bayesian network is fixed.

   Given the gene expression data D=$\{\mathbf{X}[1],...,\mathbf{X}[M]\}$, where $\mathbf{X}[m]$ (a vector of size N) consists of expression values on all N genes in the m'th sample, your goal is to estimate the parameters $\theta_{\mathbf{X_1}|\mathbf{Pa_1}},...,\theta_{\mathbf{X_N}|\mathbf{Pa_N}}$ of the Bayesian network by using maximum likelihood estimation (MLE).

   (a) **[10 points]** You decided to use table CPDs to represent statistical dependence between X_i and its parents $\mathbf{Pa_i}$. Then, for the following network (with N=3) including X_1 and its parents X_2 and X_3, describe how $\mathbf{Pa_1}$ and $\theta_{x_1|\mathbf{pa_1}}$, determine the distribution over X_1. (Hint: Use the conditional probability table we discussed in class.)



   (b) **[10 points]** Write down the likelihood function $L(\theta : D)$. Show how the decomposition of the global problem to independent sub-problems allows us to devise efficient solutions to the MLE problem. (Hint: See slide 11 of the lecture note # 2.)

   (c) **[20 points]** Prove that the MLE solution of the parameters are:

$$\hat{\theta}_{x_i|\mathbf{pa_i}} = P(X_i = x_i|\mathbf{Pa_i} = \mathbf{pa_i}) = \frac{\mathbf{M}[\mathbf{x_i}, \mathbf{pa_i}]}{\mathbf{M}[\mathbf{pa_i}]} \quad \text{for i=1,...,N} \tag{1}$$

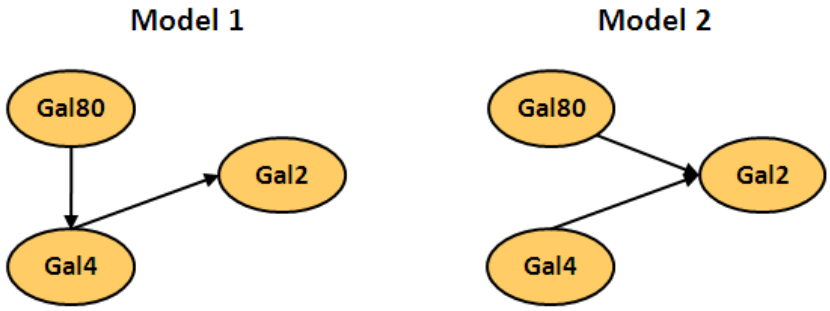
   Here, M[x_i,pa_i] means the number of samples in which $X_i = x_i$ and $\mathbf{Pa_i} = \mathbf{pa_i}$, and M[$pa_i$] means the number of samples in which $\mathbf{Pa_i} = \mathbf{pa_i}$. (Hint: Use the fact that the conditional probability is legal, i.e., $\theta_{high|\mathbf{pa_i}} + \theta_{low|\mathbf{pa_i}} = 1$.)

   i. **[5 points]** Take the log of the likelihood function in part (b).

   ii. **[15 points]** Take the derivative of the log-likelihood function in part i. with respect to the parameters θ's and find the θ's that make it zero.

2. **[60 points] Model selection to find the best regulatory network**

   In this question, we will implement an algorithm for differentiating among various structures of the regulatory network. Specifically, we will focus on two hypotheses of the galactose regulatory network in *S. cerevisiae*.



Let's assume that expression levels are binary values (high, low), and we use table CPDs for both networks in Model 1 and 2.

(a) **[10 points]** Define the parameters are in each model.

(b) **[10 points]** Given the gene expression data D measuring the binary expression levels of the 3 genes (Gal80, Gal4 and Gal2) across 112 samples, write down the likelihood function $L(\theta : D)$ for Model 1 and 2.

(c) **[10 points]** Use part (c) in Question 1 and describe how to compute the maximum likelihood estimation of the parameters in Model 1 and 2.

(d) **[25 points]** Download the data from `http://www.cs.washington.edu/homes/suinlee/genome541/disc-gal80-gal4-gal2.txt`, and implement the code that computes the likelihood score for Model 1 and Model 2. Please submit the code and the resulting scores of Model 1 and 2.

(e) **[5 points]** Select between model 1 and 2 based on the result in part (d).