CSE 527, Fall 2017
Problem Set #1
(Due Oct 23th 11:59pm)

---

1. **[15 points] Probability review**

   (a) **[5 points]** After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

   (b) **[10 points]** It is quite often useful to consider the effect of some specific propositions in the context of some general background evidence that remains fixed, rather than in the complete absence of information. The following questions ask you to prove more general versions of the product rule and Bayes' rule, with respect to some background evidence E.

   - Prove the conditionalized version of the general product rule:

   $$P(A, B|E) = P(A|B, E)P(B|E)$$

   - Prove the conditionalized version of Bayes' rule:

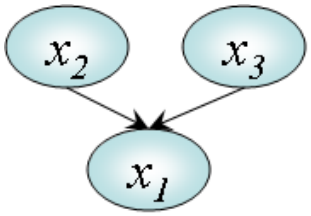
   $$P(A|B, E) = \frac{P(B|A, E)P(A|E)}{P(B|E)}$$

2. **[30 points] MLE for Bayesian networks**

   Consider N genes that are known to be involved in lipid biosynthesis. Say that we know roughly how they regulate each other's expression. However, in order to understand the molecular level mechanism of lipid synthesis or of related phenotypes (such as obesity), we need a more detailed picture of their regulatory interactions. Thus, you decided to represent these interactions by using the Bayesian network and learn its parameters from microarray expression data.

   Denote by $x_1, ..., x_N$ the variables representing the expression levels of those N genes. For simplicity, we assume that each x_i's value is discretized to two levels, i.e., $x_i \in \{up, down\}$. Let's assume that the conditional independence assumptions are already known based on the prior knowledge; so the structure of the Bayesian network is fixed. Given the data $D = \{x[1], ..., x[M]\}$, where $x[m]$ consists of an instantiation of all the variables $x_1, ..., x_N$, your goal is to learn the parameters $\theta_{x_1|pa_1}, ..., \theta_{x_N|pa_N}$ by using MLE.

   (a) **[5 points]** You decided to use the table CPDs to represent the statistical dependencies between x_i and its parents $\mathbf{pa}_i$. Then, for the following sub-network including x_1 and its parents x_2 and x_3, describe how $\mathbf{pa}_i$ and the parameters $\theta_{x_1|\mathbf{pa}_1}$, determine the distribution over x_1. (Hint: conditional distribution)



(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.

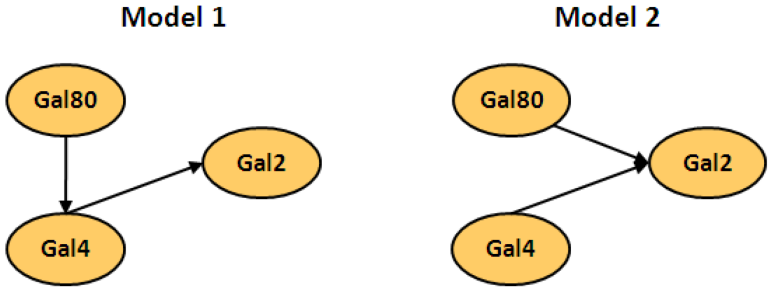
(c) **[15 points]** Prove that the MLE solution of the parameters are:

$$\hat{\theta}_{x_i|\mathbf{pa}_i} = \frac{M[x_i, \mathbf{pa}_i]}{M[\mathbf{pa}_i]} \quad \text{for } i = 1, ..., N \tag{1}$$

(Hint: Use the fact that the conditional probability is legal, i.e., $\sum_{x_i} \theta_{x_i|\mathbf{pa}_i} = 1$.)

3. **[55 points] Model selection to find the best regulatory network**

In this question, we will implement an algorithm for selecting among various structures of the regulatory network. Specifically, we will focus on two possible models 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]** What are the parameters in each model?

(b) **[10 points]** Say that we are given the gene expression data D measuring 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]** Describe how to compute the maximum likelihood estimation of the parameters in Model 1 and 2.

(d) **[20 points]** Download the data from `https://sites.google.com/a/cs.washington.edu/cse527-au17/datasets`, 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 results in part (d).