CS281A/Stat241A Homework Assignment 2 (due February 27, 2007)

1. **(ELIMINATE algorithm)**
   Suppose that we have a directed graphical model, with an arbitrary DAG $G = (V, \mathcal{E})$, and an evidence set $E \subset V$. We wish to compute a maximum *a posteriori* probability configuration of the values of variables in a subset $F \subset V$. That is, we wish to find a vector x_F^* such that

$$p(x_F^* | x_E) = \max_{x_F} p(x_F | x_E).$$

   (a) Describe (in pseudocode) a version of the ELIMINATE algorithm that is suitable for this problem. (Assume that the local conditionals are all supported on a finite set. Assume also that algorithms described in the text, such as ELIMINATE, are available as subroutines.)

   (b) Give an upper bound on the time complexity of your algorithm.

   (c) Show that your algorithm is correct.
   You can assume that ELIMINATE and MAP-ELIMINATE are correct. In particular, you need to show that the configuration returned by your algorithm is a maximizing one, and not just a sequence of members of maximizing configurations.

   (Make sure that the algorithm you propose will work correctly for the simple example of Figure 4.16 in the text. Also make sure that your algorithm's execution time is not exponential in $|F|$, in general. Notice that the MAX-PRODUCT algorithm in Figure 4.19 in the text is not suitable, since it is not appropriate for an arbitrary DAG.)

2. **(Factor graphs and polytrees)**
   Recall that the factor graph associated with a directed graph has one factor for each local conditional defined on the graph. Similarly, the factor graph associated with an undirected graph has one factor for each potential defined on the graph. (To ensure that the undirected graph completely specifies the associated factor graph, let us assume that there are no potentials associated with non-maximal cliques.)

   (a) Let G be a polytree, and let G_M be its moral graph. Let F denote the factor graph associated with G, and let F_M denote the factor graph associated with G_M. For every vertex i in G with no parents, add a factor f_i to F_M that is connected to the variable node i. Prove that F and F_M are identical. i.e., the factor graph associated with the moral graph of a polytree is the same as the factor graph associated with the polytree, modulo the single-variable factors.
   (Hint: Use induction. Work through the nodes in a topological ordering, building G_M, F_M and F.)

   (b) Prove that the factor graph associated with a polytree is a factor tree.

3. **(Naive Bayes)**
   In a pattern classification problem, a binary label $Y \in \{0, 1\}$ is to be predicted from the covariates $X_1, \ldots, X_d \in \{0, 1\}$. A *naive Bayes* model assumes that, given the class label Y, the components X_i are conditionally independent.

   (a) Specify a directed graphical model corresponding to the naive Bayes model.

   (b) Express the posterior class probability, $p(Y = 1|x)$, in terms of the prior class probability $p(Y = 1)$ and the class conditionals, $p(x_i|y)$.

   (c) Suppose we wish to use a naive Bayes to classify web pages into two classes, and let each X_w be the indicator function of the presence of word w on the page. Explain why this might not be an accurate model of the joint distribution.

   (d) Suppose we wish to make a prediction $\hat{y} \in \{0, 1\}$. It is easy to show that predicting $\hat{y} = 1$ iff $p(Y = 1|x) \geq 1/2$ minimizes $p(Y \neq \hat{y})$. Show that making this prediction using the posterior class probability for a naive Bayes model corresponds to a linear classifier, for which $\hat{y} = 1$ iff

$$\sum_{i=1}^{d} a_i X_i \geq b$$

for some real numbers $a_1, \dots, a_d, b$.

4. **(LMS algorithm)** On the course website, there is a data set (hw2.data), consisting of 30 pairs, $(x_1, y_1), \dots, (x_{30}, y_{30})$. Each x_i is a vector in $\mathbb{R}^2$, and line i of the file contains the two components of x_i, followed by y_i. We wish to use this data to estimate the parameters of a linear regression model,

$$y = \theta^T x + \epsilon,$$

where $x, \theta \in \mathbb{R}^d$ and ϵ is a zero mean Gaussian.

(a) Calculate the solution θ^* to the normal equations,

$$X^T X \theta = X^T y,$$

where X consists of the row vectors x_i^T and y is the vector of y_is.

(b) Compute the eigenvectors and eigenvalues of $X^T X$, and plot contours of the cost function $J(\theta) = (y - X\theta)^T (y - X\theta)$ in the parameter space $\mathbb{R}^2$.

(c) Plot the path through parameter space taken by the LMS algorithm when the initial parameter value is 0. Use three values of the stepsize:

i. $\rho = 2/\lambda_{max}$,

ii. $\rho = 1/(2\lambda_{max})$,

iii. $\rho = 1/(8\lambda_{max})$,

where λ_{max} is the largest eigenvalue of $X^T X$.