CS281A/Stat241A Homework Assignment 2 (due 5pm September 30, 2009)

1. **(Polynomial representation)**
Consider an undirected graphical model with potentials $\psi_C(x_C)$ defined for each C in the set $\mathcal{C}$ of maximal cliques. Suppose that the random variables $X_1, \ldots, X_n$ are discrete, and let $\mathcal{X}_i$ denote the set of values that X_i can take. We can think of this class of probability distributions as parameterized by the real numbers $\psi_C(x_C)$, where C ranges over the set of maximal cliques and x_C ranges over all values of the variables in C. In this question, we investigate the representation of a joint probability as a function of these parameters.

   Fix a set $S \subset \{1, \ldots, n\}$, and an assignment y_S to the variables in S.

   (a) Show that we can construct a function $f(\theta, \delta)$ in the variables $\{\theta_{C,x} : C \in \mathcal{C}, x \in \mathcal{X}_C\}$ and $\{\delta_{i,x} : i = 1, \ldots, n, x \in \mathcal{X}_i\}$ so that f is poynomial in δ and we can write

   $$p(X_S = y_S) = f(\theta^\psi, \delta^{y_S}),$$

   where for each assignment x_C to the variables in C,

   $$\theta^\psi_{C,x_C} = \psi_C(x_C),$$

   and for each assignment $x_i \in \mathcal{X}_i$ to the random variable X_i,

   $$\delta^{y_S}_{i,x_i} = \begin{cases} 0 & \text{if } i \in S \text{ and } x_i \neq y_i, \\ 1 & \text{otherwise.} \end{cases}$$

   (b) Show how we can compute $p(X_F | X_E = \bar{x}_E)$ in terms of the polynomial f.

   (c) For some $i \in S$, let y'_S be an assignment to the variables in S that satisfies $y'_j = y_j$ for $j \in S \setminus \{i\}$. Show that

   $$p(X_S = y'_S) = \frac{\partial f(\theta^\psi, \delta^{y_S})}{\partial \delta_{i,y'_i}}.$$

   (d) Show that when we remove an observation of a variable y_i, the probability becomes

   $$p(X_{S\setminus\{i\}} = y_{S\setminus\{i\}}) = \sum_{y'_i \in \mathcal{X}_i} \frac{\partial f(\theta^\psi, \delta^{y_S})}{\partial \delta_{i,y'_i}}.$$

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. (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 $a_1, \ldots, a_d, b$.

4. **(LMS algorithm)** On the course website, there is a data set (hw2.data), consisting of 30 pairs, $(x_1, y_1), \ldots, (x_{30}, y_{30})$. 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 $X^T X$.