CS281A/Stat241A Homework Assignment 3 (due 5pm October 14, 2009)

1. **(Logistic regression)**
   Suppose that we wish to model an unordered discrete response variable Y (such as the outcome of a potential customer's visit to a website), conditioned on a vector X of real variables (such as characteristics of the advertisements presented to the potential customer). We could model this kind of relationship using
   $$\Pr(Y = y|X = x) = \frac{\exp(-\beta_y' x)}{1 + \sum_{i=1}^{k-1} \exp(-\beta_i' x)}, \tag{1}$$
   where $y \in \{1, \ldots, k-1\}$, $x \in \mathbb{R}^d$, $\beta_i \in \mathbb{R}^d$ and k is the number of distinct responses.

   Suppose that we have data $(x_1, y_1), \ldots, (x_n, y_n)$ generated i.i.d. from the model (1).

   (a) Write down the log likelihood and its first and second derivatives.

   (b) Describe (in pseudocode) a Newton-Raphson algorithm for maximizing the log likelihood.

   (c) Suppose that the dimension d of the data is so large that it is impractical to store more than a constant number of vectors in $\mathbb{R}^d$, let alone manipulate second derivative matrices. Suggest an online steepest ascent algorithm.

2. **(ML Estimation)**
   On the course website, there is a data set (hw3-2.data), consisting of 100 pairs, $(v_1, y_1), \ldots, (v_{100}, y_{100})$. Each v_i is a vector in $\mathbb{R}^2$, and each y_i is a number in $\{1, \ldots, 4\}$. Line i of the file contains the two components of v_i, followed by y_i. Using the algorithm that you proposed in question 1b, calculate the maximum likelihood estimate for the parameters of the model (1) for this data, with $x_i = (1, v_{i1}, v_{i2})'$. Plot the data (with four different symbols for the y values) and the contours
   $$C_y = \left\{ v \in \mathbb{R}^2 : \Pr\left(Y = y \,\middle|\, X = \begin{pmatrix} 1 \\ v \end{pmatrix} \right) = 1/2 \right\}$$
   for $y = \{1, \ldots, 4\}$.

3. **(IPF)** Consider the undirected graphical model
   $$p(x) = \frac{1}{Z} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j),$$
   with binary variables $x_1, \ldots, x_k$, where the $\psi_{i,j}$ are non-negative functions. The data in the file `hw3-3.data` on the course website consists of n binary vectors of length $k = 5$. Implement the IPF algorithm, and use your implementation on this data to estimate the model parameters for the following graphs:

   (a) $E = \{(1,2), (2,3), (3,4), (4,5), (5,1)\}$,

   (b) $E = \{(1,2), (2,3), (3,4), (4,1), (2,5)\}$,

   (c) $E = \{(1,2), (2,3), (3,4), (2,5)\}$.

   Which model fits the data best?