CS281A/Stat241A Homework Assignment 3 (due October 15, 2015)

1. **(Multiclass Classification)**
   Suppose that we wish to model an unordered discrete response variable Y (such as which folder should be suggested as the destination for a document), conditioned on a vector X of real variables (such as features of words in the document).

   (a) *(Logistic regression)* We could model this kind of relationship using

   $$\Pr(Y = y|X = x) = \frac{\exp(-\beta_y^\top x)}{1 + \sum_{i=1}^{k-1} \exp(-\beta_i^\top x)}, \tag{1}$$

   where $y \in \{1, \ldots, k-1\}$, $x \in \mathbb{R}^d$, $\beta_y \in \mathbb{R}^d$ 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).

      i. Write down the log likelihood and its first and second derivatives.
      ii. Describe (in pseudocode) a Newton-Raphson algorithm for maximizing the log likelihood.

   (b) *(Linear regression)* We might also approach multiclass classification through a general linear model. This is a generalization of linear regression in which the response variable is in $\mathbb{R}^k$; thus, we have $\beta \in \mathbb{R}^{d \times k}$. The probability model is

   $$p(y|X = x, \mathbf{\Sigma}, \beta) = (2\pi)^{-\frac{k}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \beta^\top x)^\top \mathbf{\Sigma}^{-1}(y - \beta^\top x)\right\}. \tag{2}$$

   Given data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}^k$, derive the maximum likelihood estimate of β in the spherical case, where $\mathbf{\Sigma} = \boldsymbol{I}$.

2. **(Comparison of multiclass classifiers)** On the course website, there is a data set (hw3-2-train.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. Use the covariates $x_i = (1, v_{i1}, v_{i2})^\top$.

   (a) Implement the algorithm that you proposed in Question 1a, and use it to calculate the maximum likelihood estimate $\hat{\beta}$ for the parameters of the model (1) for this data.

   (b) Given the data $(x_1, y_1), \ldots, (x_{100}, y_{100}) \in \mathbb{R}^3 \times \{1, \ldots, 4\}$, we can use the response variables $\tilde{y}_i = e_{y_i}$ (where e_i is the ith standard basis vector in $\mathbb{R}^4$). Implement the algorithm that you proposed in Question 1b, and use it to calculate the maximum likelihood estimate $\hat{\beta}$ for the parameters of the model (2) for the data $(x_1, \tilde{y}_1), \ldots, (x_{100}, \tilde{y}_{100})$.

   In both cases:

      (i) Use the parameter estimates to predict the labels for the training data (in hw3-2-train.data) and the test data (in hw3-2-test.data, also on the course website), by predicting the index y that maximizes $p(y|x, \hat{\beta})$. Report the training and test misclassification rates.
      (ii) Plot the training data (with four different symbols for the y values) and the boundaries between the regions $C_1, \ldots, C_4$, where

      $$C_y = \left\{v \in \mathbb{R}^2 : p(y|x, \hat{\beta}) > \max_{y' \neq y} p(y'|x, \hat{\beta})\right\}.$$

   (c) Why is logistic regression substantially better?

3. **(Exponential Families)**
Recall that a set of probability distributions is an exponential family if it takes on the following form:

$$\left\{x \mapsto p(x;\eta) = h(x)\exp\left(\eta^\top T(x) A(\eta)\right) : \eta \in \Theta\right\},$$

for the natural parameter η, sufficient statistic T, log-normalization A, and reference measure h. Determine if each of the following sets of distributions is an exponential family. For those that are not, give a maximal subset that is an exponential family. In each case, write down η, T, A, and h.

(a) Geometric: parameters $p > 0$, support $k = \{1, 2, \ldots\}$

$$p(k;p) = (1-p)^k\, p.$$

(b) Pareto: Parameters $x_m > 0$ and $\alpha > 0$

$$p(x; x_m, \alpha) = \frac{\alpha\, x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} 1[x \geq x_m].$$

(c) Inverse Gaussian: parameters $\lambda > 0$, $\mu > 0$

$$p(x; \lambda, \mu) = \left[\frac{\lambda}{2\pi x^3}\right]^{1/2} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x}.$$

(d) Weibull: parameters $\lambda > 0$, $k > 0$

$$p(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

4. **(Deviation inequalities)**
For a random variable X with support in the interval $[a, b]$ and density f, consider the exponential family

$$\begin{aligned} &\{p(x) = f(x)\exp(\eta x - A(\eta)) : \eta \in \mathbb{R}\}, \\ &A(\eta) = \log \mathbb{E} e^{\eta X}. \end{aligned}$$

In this question, we use properties of this exponential family to prove a deviation inequality for X.

(a) Show that, for any $t > 0$ and any $\eta > 0$,

$$\log \Pr(X \geq t) \leq A(\eta) - \eta t.$$

(b) Hence show that, for $\epsilon > 0$,

$$\Pr(X \geq \mathbb{E}X + \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right).$$

(Hint: consider the second-order Taylor expansion of A about 0:

$$A(\eta) = A(0) + \eta \nabla A(0) + \frac{\eta^2}{2}\nabla^2 A(\xi),$$

for some $\xi \in [0, \eta]$.)