CS 281A/Stat 241A Homework Assignment 3 (due October 16)

1. **Dirichlet expectations.**

   The *Dirichlet distribution* is a continuous distribution on the K-simplex, $\{\theta = (\theta_1, \theta_2, \ldots, \theta_K), \text{such that } \theta_i \geq 0 \text{ for } i = 1, \ldots, K, \text{ and } \sum_{i=1}^K \theta_K = 1\}$:

   $$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1},$$

   where $\alpha_i > 0$ are parameters.

   (a) Compute $\mathrm{E}[\theta_k]$. [Hint: do it directly.]

   (b) Compute $\mathrm{Cov}[\theta_j \theta_k]$. [Hint: do it directly.]

   (c) Compute $\mathrm{E}[\log \theta_k]$. [Hint: Show that the Dirichlet distribution is in the exponential family and take a first derivative of the cumulant function.]

2. **Dirichlet-multinomial prediction.**

   Let $\theta \sim \mathrm{Dir}(\alpha)$. Consider multinomial random variables $(X_1, X_2, \ldots, X_N)$, where $X_n \sim \mathrm{Mult}(\theta)$ for each n, and where the X_n are assumed conditionally independent given θ. Now consider a random variable $X_{\text{new}} \sim \mathrm{Mult}(\theta)$ that is assumed conditionally independent of $(X_1, X_2, \ldots, X_N)$ given θ. Compute:

   $$p(x_{\text{new}} \mid x_1, x_2, \ldots, x_N, \alpha)$$

   by integrating over θ. [Hint: Your result should take the form of a ratio of gamma functions.]

3. **Conjugacy and prediction.**

   Redo the preceding problem, replacing the multinomial distribution with an arbitrary exponential family distribution, and the Dirichlet distribution with the corresponding exponential family conjugate distribution. You are to show that in general the predictive probability $p(x_{\text{new}} \mid x_1, x_2, \ldots, x_N)$ is a ratio of normalizers.

4. **Bayesian inference in a multinomial directed tree**.

   Consider a directed graphical model with n nodes, $X = (X_1, \ldots, X_n)$. Assume that the graph is a tree, i.e., each node i except for the root has exactly one parent $\pi(i)$. Each node X_i takes on a value in $\{1, \ldots, K\}$. We assume that the conditional probabilities have the following parameterization:

   $$p(X_i = b \mid X_{\pi(i)} = a) = \theta_{ab}, \quad a, b \in \{1, \ldots, K\}.$$

   The parameters of the model are therefore

   $$\theta = \{\theta_{ab} : 1 \leq a, b \leq K\}$$

   with $\sum_{b=1}^K \theta_{ab} = 1$ for all $1 \leq a \leq K$. This probability model is written succinctly as $p(x \mid \theta)$.

   (a) What are the sufficient statistics $T(X)$ for the model $p(x \mid \theta)$?

   (b) What is the maximum likelihood estimate for θ?

   (c) Give a conjugate prior $p(\theta)$.

   (d) Based on that conjugate prior, what is the posterior $p(\theta \mid x)$?

5. **Classification**.

   The course website contains a data set "classification.dat" of (x_n, y_n) pairs, where the x_n are 2-dimensional vectors and y_n is a binary label.

(a) Plot the data, using 0's and X's for the two classes. The plots in the following parts should be plotted on top of this plot.

(b) Fit a generative model to the data, using Gaussian class-conditional densities with equal covariance matrices. Calculate the posterior probability of class 1, and plot the line where this probability is equal to 0.5.

(c) Write a program to fit a logistic regression model using the IRLS algorithm (remembering to include the intercept term). Plot the line where the logistic function is equal to 0.5.

(d) Write a program to fit a logistic regression model using stochastic gradient ascent. Plot the line where the logistic function is equal to 0.5.

(e) Fit a linear regression to the problem, treating the class labels as real values 0 and 1. (You can solve the linear regression in any way you'd like, including solving the normal equations, using the LMS algorithm, or calling the built-in routines in Matlab or Splus). Plot the line where the linear regression function is equal to 0.5.

(f) The data set "classification.test" is a separate data set generated from the same source. Test your fits from parts (b), (c), (d) and (e) on these data and compare the results.