B555: Homework 3

1. Let S be the collection of all of the binary strings no more than 4 digits long including the empty string e. Associate each string, $s \in S$ with a random variable, X_s such that

$$\begin{aligned} X_{s\circ 1} &= X_s + Y_{s\circ 1} \\ X_{s\circ 0} &= X_s + Y_{s\circ 0} \end{aligned}$$

where $\circ$ denotes concatenation and X_e and $\{Y_s : s \in S\}$ are all mutually independent $N(0,1)$ random variables.

   (a) Draw a directed acyclic graph labeleling vertices with X_s random variables such that the graph represents the conditional independence structure of the probability distribution.

   (b) Give a factorization of the joint density on the variables $\{X_s : s \in S\}$.

   (c) What is the conditional expectation of X_e given all of the variables $\{X_s : s \in S, s \neq e\}$.

2. Suppose you observe vectors $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N(\mu, \Sigma)$ where the mean vector μ is unknown and the covariance matrix Σ is known. Find the maximum likelihood estimator for μ, $\hat{\mu}_{\text{MLE}}$.

3. Fisher's famous iris data set measures sepal width, petal width, sepal length, and petal length for three different types of iris: Virginica, Versicolor, and Setosa. This problem requires you to build a Gaussian classifier using these data. To read the data set into R use

```
data(iris)
```

   If, for instance you want to create the empirical mean and covariance matrix

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_i x_i \\ S &= \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^t \end{aligned}$$

   for the Versicolor class you could use

```
X = iris[iris[,5] == "versicolor",1:4]
mv = as.matrix(colMeans(X));
Sv = as.matrix(cov(X))
```

   where the as.matrix cast is important. (The 5th column has the class labels and the first four columns are the feature values.) You can get the inverse and determinant of S with

```
Cv = solve(Sv)
dv = det(Sv)
```

   (a) Create a scatter plot using the first two features with different plot symbols for each iris type.

   (b) Using all of the data compute the empirical mean vectors and covariance matrices for the 3 classes. Use these to classify the data vectors using an optimal Bayes' classifier assuming the class conditionals are multivariate Gaussian. Show a scatter plot of the first two feature values as before, now coloring the correctly classified vectors in green and the incorrect ones in red.

   (c) The above apporach is not completely honest since we tested on the same data we used to train our classifier. This gives the classifier an unfair advantage. Now train your classifier using the first half of the data for each of the three classes and test on the 2nd half. Show a similar plot.

(d) For both classifiers, compute the error rate using the 2nd half of the 3 data sets. Do you get similar results for the two methods. Explain why or why not.

4. In the "earthquakes.r" example (on the web page) we showed how our information about the Poisson rate, λ was updated sample-by-sample using the conjugate prior pair of

$$\lambda \sim \text{Gamma}(\theta, k)$$
$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$$

Now assume that

$$\mu \sim \text{Normal}(\nu, \rho^2)$$
$$x_1, \ldots, x_n \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$$

with $\nu = 0, \rho^2 = 1,000,000, \sigma^2 = 1$.

(a) In R, generate samples $x_1, x_2, \ldots$, from the Normal$(10, 1)$ distribution and plot the conditional pdfs $p(\mu|x_1, \ldots, x_n)$ for $n = 1 \ldots 100$.

(b) Describe how the conditional densities change as you receive more and more data.

(c) What appears to be the limit of $E[\mu|x_1, \ldots, x_n]$ as n increases?

5. $X \sim \text{Beta}(\alpha, \beta)$ if X has pdf

$$p(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $B(\alpha, \beta)$ is the constant necessary to make the pdf integrate to 1 (we don't need to know how to compute B in this problem). It may help to know that the mean of a Beta(α, β) distribution is $\frac{\alpha}{\alpha+\beta}$. Suppose $q \sim \text{Beta}(\alpha, \beta)$ and $X \sim \text{Binomial}(n, q)$.

(a) What is the the conditional density $p(q|x)$.

(b) If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Binomial}(n, q)$ what is $p(q|x_1, \ldots, x_n)$?

(c) Suppose we want to estimate q with some number $\hat{q}$. if our estimate misses the true value of q by distance d we will pay $\$d^2$. How should $\hat{q}$ be chosen so that we minimize our expected cost?

6. Consider the classifier that classifies according to

$$\hat{c} = \arg\max_c p(c|x)$$

where c is the class and x is the data. What is the loss function so that $\hat{c}$ minimizes expected loss?

7. In this problem we perform logistic regression using Fisher's iris data. While R provides off-the-shelf methods for doing this, here you are to construct everything by hand.

(a) Read in the data and create a class vector that is 1 for the *setosa* iris and 0 otherwise.

(b) Using all four variables as features, initialize your weight vector to be all 0's and perform gradient ascent to optimize the data likelihood. You may need to experiment with step size to make sure your algorithm is stable. Every iteration in your gradient ascent algorithm should print out the log likelihood of the data to verify that this increases monotonically.

(c) Using your resulting weight vector classify all of the iris flowers as either *setosa* or not by choosing the class with the greatest posterior probability according to your trained model. You should be able to classify all of these correctly.