B555: Midterm Exam

1. Suppose that a random vector (x_1, x_2, x_3) has probability density function

$$f(x_1, x_2, x_3) = ke^{-(3x_1^2+2x_2^2+2x_3^2+x_1x_2+x_2x_3)/2}$$

where $x_1, x_2, x_3 \in \Re$ and k is a constant.

   (a) (3 pts) Give an expression for k without evaluating the expression.

   (b) (12 pts) Show that the variables x_1, x_2, x_3 have a joint normal distribution for some mean vector μ and inverse covariance matrix Σ^{-1} which you will construct.

   (c) (5 pts) Draw the directed acyclic graph that represents the conditional independence structure of the variables and identify any conditional independencies.

   (d) (5 pts) Suppose the covariance matrix, Σ is given. Say how you would determine any independence relations that exist among x_1, x_2, x_3.

2. Consider a two-class classification problem where we try to predict the class, $C \in \{1, 2\}$ of a random data vector X using logistic regression.

   (a) (5 pts) In logisitic regression write the parametric form we assume of $P(C = 1|X = x)$

   (b) (10 pts) Assume that our data are given by an $n \times k$ data matrix and $n \times 1$ class vector

$$\begin{pmatrix} x_{11} & \dots & x_{1k} \\ x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

   In estimating the k-dimensional weight vector, w, write the objective function we seek to maximize, explicitly showing the dependence on the weight vector and the data given above.

   (c) (7 pts) Suppose that a component of the weight vector, w_j, is 0 in the logistic regression model. Writing C for the class random variable and $X = (X_1, \dots, X_k)$ for the random vector of predictors, interpret this fact as a statement of independence or conditional indepdendence between the variables $C, X_1, \dots, X_k$.

3. Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(p). That is, the $\{X_i\}_{i=1}^n$ are independent and $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Let $Y = X_1 + \ldots + X_n$.

   (a) (5 pts) What is the probability mass function for Y?

   (b) (5 pts) What are the mean and variance for Y. It may help to use the definition of Y given above.

   (c) (10 pts) Show that the maximum likelihood estimator for p is the sample mean.

   (d) (10 pts) Suppose we want to guard against estimating p as 0 or 1 no matter what the data indicate. We do this by "padding" our data with m negative observations and m positive observations. That is, we imagine our data to be

$$\underbrace{0, 0, \dots, 0}_{m \text{ times}} \underbrace{1, 1, \dots, 1}_{m \text{ times}} x_1, x_2, \dots, x_n$$

   and proceed using the sample mean of the padded data as the estimate of p. Using the Beta(α, β) distribution as your prior distribution, what are the parameters of the Beta, and the estimation strategy (MLE, MSE, MAP, etc) so that your estimate is the same as the padded sample mean. It may help to recall that Beta distribution has density function

$$p(x|\alpha, \beta) = kx^{\alpha-1}(1-x)^{\beta-1} \quad x \in (0, 1).$$

   and that the mean of the Beta is $\frac{\alpha}{\alpha+\beta}$.

4. Suppose $P\%$ of the cases appearing in a certain justice system involve defendants who are guilty of the crime for which they are accused. The judicial process counts the incriminiating factors as evidence. We know that this number has Poisson distribution

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

for $x = 0, 1, \ldots$. We believe that $\lambda = 5$ when the person is guilty, while $\lambda = 1$ when the person is innocent.

   (a) (10 pts) Give an expression for the probability of guilt given the assumptions above as a function of x.

   (b) (7 pts) Suppose the justice system hears one case a day for many years and seeks to minimize the number of incorrect decisions that are made on average. Suppose that the system convicts when $x > x^*$. Explain in detail how to compute x^*, involving computer code in R if you cannot express the answer in closed form.

   (c) (7 pts) Suppose we believe that an incorrect conviction is 100 times as bad as an incorrect acquittal. That is, the cost of an incorrect conviction is $100C$ where C is the cost of an incrrect acquittal. What would the expected cost be if we convict when $x > x^*$?

5. Suppose we have a collection of observations $(x_1, y_1), \ldots, (x_n, y_n)$ where the $\{x_i\}$ and $\{y_i\}$ are real valued. We want to estimate y by

$$\hat{y}_i = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3$$

for some coefficients w_0, w_1, w_2, w_3. We will choose the coefficients so that $\sum_i (y_i - \hat{y}_i)^2$ is minimized.

   (a) (10 pts) Give an expression for the optimal coefficients, w_0, w_1, w_2, w_3 as a function of $(x_1, y_1), \ldots, (x_n, y_n)$. Be sure to define precisely any matrices or vectors you use.