B555: Final Exam

1. Suppose we choose a class $c \in \{1, 2\}$ according to $P(c = k) = \pi_k$, and then choose a D-dimensional vector $x \sim N(\mu_k, \Sigma_k)$ when k is the class.

   (a) (5 pts) What is the probability density function for x according to this model?

   (b) (10 pts) Suppose we observe a random sample from this model: $(x_1, c_1), (x_2, c_2), \ldots, (x_N, c_N)$. Give the MLEs for $\pi_1, \pi_2, \mu_1, \mu_2$, and Σ_1, Σ_2.

   (c) (5 pts) Suppose the model parameters are known and you observe a new vector x. How would you classify x so as to minimize the probability of misclassification?

2. Suppose we observe D-dimensional predictor vectors $x_1, \ldots, x_N$ and real-valued response variables $y_1, \ldots, y_N$ and assume $y_n = w^t x_n + e_n$ where the $e_1, \ldots, e_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Let X and y be our usual data matrix and response vector

$$X = \begin{pmatrix} \ldots & x_1 & \ldots \\ \ldots & x_2 & \ldots \\ \vdots & \vdots & \vdots \\ \ldots & x_N & \ldots \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

   (a) (5 pts) We wish to construct the globally best collection of three predictor variables. That is, if $X(j_1, j_2, j_3)$ is the $N \times 3$ matrix constructed from columns j_1, j_2, j_3, we wish to find

$$\min_{j_1, j_2, j_3, w} ||y - X(j_1, j_2, j_3)w||^2$$

   Describe in detail how to find the *globally* optimal j_1, j_2, j_3 and associated w?

   (b) (10 pts) Now suppose we want to choose k predictor variables. Describe precisely the forward variable selection method for constructing the k variables $j_1, \ldots, j_k$

   (c) (5 pts) Suppose that $j_1, j_2, \ldots$ are the variables chosen by the forward selection method and that,

$$w_k = \arg\min_w ||y - X(j_1, \ldots, j_k)w||^2$$

   That is, w_k is the optimal weight function using the first k variables from forward selection. Graph a possible result for

$$V(k) = ||y - X(j_1, \ldots, j_k)w_k||^2$$

   for $k = 1, \ldots, D$.

   (d) (5 pts) Suppose our training set has $N = D = 100$ and $(x'_1, y'_1), \ldots, (x'_N, y'_N)$ are an independent validation set following the same model as $(x_1, y_1), \ldots, (x_N, y_N)$. Graph a characteristic result for

$$V'(k) = ||y' - X'(j_1, \ldots, j_k)w_k||^2$$

   for $k = 1, \ldots, D$, where y' and X' are the response vector and data matrix for the validation set and the $\{j_k\}$ and $\{w_k\}$ are those chosen in the forward selection procedure. In other words, what happens when we apply the variables and weight vectors from forward selection to new data.

3. Suppose we have a training set $(x_1, t_1), \ldots, (x_N, t_N)$ where $x_n \in \Re^D, t_n \in \{1, -1\}$ for $n = 1, \ldots, N$. We form the function $y(x) = w^t x + b$ and classify each point x_n as

$$\text{sgn}(y(x_n)) = \text{sgn}(w^t x_n + b)$$

   where sgn(x) is 1/-1 as x is positive/negative. Suppose we adopt the strategy of the SVM and seek a classifier that maximizes the margin over the training data. Assume we insist on classifying all of the training points correctly.

   (a) (5 pts) Express the optimal w, b directly as the values that maximize the margin over the training data.

   (b) (5 pts) Express the optimal w, b as the solution to a quadratic programming problem.

   (c) (5 pts) Can there be two different solutions to your optimization problem giving the same maximal margin? Either show an example where there are two such optimal solutions or argue that this cannot occur.

4. Consider the 3-class logistic regression problem on a D-dimensional feature vector, x.

   (a) (5 pts) What parametric form does the logistic regression model assume for $P(\text{Class} = k|x)$?

(b) (5 pts) Suppose our data are $(x_1, c_1), \ldots, (x_N, c_N)$ where $x_n \in \Re^D$ and $c_n \in \{1, 2, 3\}$. In this model we arrive at the estimated weights, $\hat{w}_1, \hat{w}_2, \hat{w}_3$, by optimizing what objective function?

(c) (5 pts) Given a generic x, and $\hat{w}_1, \hat{w}_2, \hat{w}_3$, how would you classify x?

(d) (5 pts) Suppose the loss for classifying class k as class k' is $|k - k'|$. Now how would you classify x?

5. Suppose there are two dice, one fair and one biased. In a set of N repeated experiments one of the dice is chosen at random, rolled D times, and we observe the results: $x_{n1}, \ldots, x_{nD}$ where $x_{nd} \in \{1, \ldots, 6\}$. We suppose that the probabilility of choosing the fair die is π_0 while $\pi_1 = 1 - \pi_0$ is the probability of choosing the biased die. These probabilities are unknown to us. The biased die gives the faces $1, \ldots 6$ with probabilites $p_1, \ldots, p_1$.

(a) (5 pts) Letting X be the $N \times D$ data matrix, write down the data log likelihood $\log P(X|\pi_0, \pi_1, p_1, \ldots, p_6)$

(b) (5 pts) What is

$$\gamma_n = P(\text{fair die}|x_{n1}, \ldots, x_{nD})?$$

(c) (10 pts) Suppose we have current estimates $\pi_0^{\text{old}}, \pi_1^{\text{old}}, p_1^{\text{old}}, \ldots, p_6^{\text{old}}$. Describe in detail how the EM algorithm would compute $\pi_0^{\text{new}}, \pi_1^{\text{new}}, p_1^{\text{new}}, \ldots, p_6^{\text{new}}$

6. (10 pts) Suppose $p(x_{n+1}|x_1, \ldots, x_n)$ does not depend on $x_1, \ldots, x_{n-1}$. That is, the function $p(x_{n+1}|x_1, \ldots, x_n)$ does not change as different values are plugged in for $x_1, \ldots, x_{n-1}$ Using the definition of conditional probability show

$$p(x_{n+1}|x_1, \ldots, x_n) = p(x_{n+1}|x_n).$$