B555: Midterm 2 (110 pts)

1. Suppose we have collection of d-dimensional observations $x_1, \ldots, x_n$ and would like to estimate a Gaussian mixture model, $p(x)$, with K mixture components. For notation we let $\pi_1, \ldots, \pi_K$ be the probabilities of the K mixture components, and let μ_k,Σ_k $k = 1, \ldots, K$ be the mean and covariance for the kth component.

   (a) (5 pts) Give the parametric form of $p(x)$ according to the GMM.

   (b) (5 pts) Suppose that we want to use the EM algorithm to estimate the GMM model and let $\pi_k^{\text{old}}, \mu_k^{\text{old}}, \Sigma_k^{\text{old}}$ be the current parameter estimates. To do this we assume unobserved variables c_i, $i = 1, \ldots, n$ giving the *true* mixture component for each x_i. Show explicitly how to compute $P(c_i = k|x_i)$?

   (c) (10 pts) Give update equations for $\pi_k^{\text{new}}, \mu_k^{\text{new}}, \Sigma_k^{\text{new}}$ in terms of the $\{x_i\}$, the parameters $\{\pi_k^{\text{old}}, \mu_k^{\text{old}}, \Sigma_k^{\text{old}}\}$, and anything else you need to derive from these quantities.

2. Suppose we have data $(x_1, c_1), \ldots, (x_n, c_n)$ where each x_i has 100 binary components and each c_i is one of K possible classes. We would like build a classifier that will classify new data according to these K classes.

   (a) (5 pts) What are the quantities (parameters, distributions, etc.) that must be estimated to approximate the Bayes classifier?

   (b) (5 pts) What are the quantities that must be estimated to approximate the Naive Bayes classifier?

   (c) (5 pts) Suppose that we wish to distinguish K cities, and $x_i = x_{i,1}, \ldots, x_{i,100}$ is created by sampling (at random) 100 voters from the city and asking one of 100 different yes-or-no questions. For the jth voter it is known that we will ask the jth question — we use the questions in order. Say how you would model the class-conditional distributions, $p(x|\text{City} = k)$ where x is the vector of 100 binary answers.

3. Suppose we have a sequence of observations $(t_1, y_1), \ldots, (t_n, y_n)$, where t_i measures the time aging a particular sample of cheese and y_i measures the content of a certain compound after the aging process. We would like to build a model that estimates y as a Mth order polynomial in t:

$$\hat{y}(t) = \hat{\alpha}_M t^M + \hat{\alpha}_{M-1} t^{M-1} + \ldots + \hat{\alpha}_1 t + \hat{\alpha}_0$$

   (a) (10 pts) Explain, in detail, how to estimate the $\hat{\alpha}$ parameters using linear regression, giving an explicit formula.

   (b) (5 pts) Give the statistical model for regression that relates the y's to the t's in this case, stating the assumed distribution of all random quantities.

   (c) (5 pts) Suppose that we believe the statistical model for regression. Is the linear regression estimate, $\hat{\alpha}_0$, unbiased for the true value, α_0? Explain in detail what is meant by unbiasedness in this case.

   (d) (5 pts) Suppose that we want to estimate the model according to *ridge* regression with penalty parameter λ. Explain, in detail, how to estimate the vector, α in this case.

   (e) (5 pts) Is $\hat{\alpha}_0$ obtained from ridge regression unbiased? Explain why or why not?

4. Suppose we have $(w_1, f_1, s_1), \ldots, (w_n, f_1, s_n)$ where each w_i and f_i measure the amount of water and fertilizer a given tomato plant receives, and s_i measures the sugar content of the plant. Suppose we want to create a regression model that tries to predict future values of s from observed (w, f) values. We would like to do this by generalizing the idea of classification trees to the regression scenario. A natural impurity measure for a collection of real values $s_1, \ldots, s_n$ is

$$I(s_1, \ldots, s_n) = \sum_i (s_i - \bar{s})^2$$

   where $\bar{s} = \frac{1}{n} \sum_i s_i$ is the sample mean. $I(s_1, \ldots, s_n)$ will be small when the s's are clustered close together and large when they are spread out.

   (a) (6 pts) Give a formula for choosing the first split of your regression tree by greedily minimizing your impurity measure.

   (b) (6 pts) Suppose that a terminal node of your regression tree is associated with the subsample of size m of the original data given by $(w_{i_1}, f_{i_1}, s_{i_1}), \ldots, (w_{i_m}, f_{i_m} s_{i_m})$. What single number should you choose to predict s if you arrive at this terminal node in your tree? Say why you would choose this number.

5. Suppose we want to perform K-class logistic regression on a d-dimensional data vector, x, in an effort to predict the class of x.

   (a) (5 pts) What is the form of $P(\text{Class} = k|x)$ according to the logistic regression model in terms of the weight vectors $w_1, \ldots, w_K$?

(b) (5 pts) Suppose our data are $(x_1, c_1), \ldots, (x_n, c_n)$ where the $\{x_i\}$ are K-dimensional vectors and the $\{c_i\}$ are the classes. Express the optimal values, $\hat{w}_1, \ldots, \hat{w}_K$ as the solution to a minimization problem.

(c) (5 pts) Suppose we solve for the weight vectors numerically using a data set of training examples, as is usual for logistic regression, and suppose $\hat{w}_1, \ldots, \hat{w}_K$ are the locally-optimal values that result. Express this local optimality in terms of an equation that the $\hat{w}_1, \ldots, \hat{w}_K$ satisfy.

(d) (5 pts) Suppose that the assumptions of logistic regression are known to be true, and we have been given the true weight vectors $w_1, \ldots, w_K$. Suppose we would like to compute the class-conditional distributions: $p(x|\text{Class} = k)$. Explain how to do this using only the weight vectors, or why, in detail, it is not possible?