1. The data file D for designing a classifier has $N_v$ patterns, each belonging to one of $N_c$ classes. Here a pattern is one example input vector $x_p$ followed by its class ID $i_c(p)$. Assume that there are equal numbers of patterns from each class and that the patterns are ordered with class 1 patterns first, class 2 patterns next, etc. J-fold validation is to be used to help perform structural risk minimization (SRM).

(a) Find $N_v(i)$, the number of patterns from the ith class.
(b) When J-fold validation is used, how many training/validation set pairs will there be ?
(c) In Leave-One-Out (LOO) validation, each of the $N_v$ patterns becomes a one-pattern validation set, with the other ($N_v - 1$) patterns being used in a training set. For what value of J is J-fold validation equivalent to the LOO method ?
(d) Let $N_{vt}(J)$ and $N_{vv}(J)$ denote the numbers of patterns in the training and validation sets, in terms of $N_v$ and J. Give expressions for $N_{vt}(J)$ and $N_{vv}(J)$.
(e) Given the ordering of the patterns in file D, what would be the worst possible method for assigning patterns to J non-overlapping sets for J-fold validation ? (Give the worst value of J). What value of $P_{ev}$ might we observe for this case ?
(f) If we want to design a good final classifier for a given application, using file D, what procedure do we use ? (How do we form training and/or validation sets and how many classifiers should be designed, etc.)

2. Given input vectors $x_p$ for training a classifier, with elements $x_p(n)$, a measure of linear dependence of the nth feature is

$$E(n) = \frac{1}{N_v} \sum_{p=1}^{N_v} [x_p(n) - \sum_{m \neq n} a_m x(m)]^2$$

(a) Given n, find a set of linear equations, in terms of an $r(n,m)$, that must be solved to minimize E(n) with respect to $a_m$. Define $r(n,m)$.
(b) In part (a), how many equations and how many unknowns do we have ?
(c) Give a modified version of $E(n)$ that takes into account the value of $\sigma_n^2 = \text{var}(x(n))$.
(d) Must the coefficients $a_m$ be found only once, or once for each value of n ?

3. A nearest neighbor classifier uses the squared Euclidean distance $d(x,m_{ik}) = [x - m_{ik}]^T [x - m_{ik}]$ where the column vector $x$ has many linearly dependent elements. Suppose that $x$ can be expressed as $x = A \cdot y$, where $y$ has no linearly dependent elements. Here, $A$ has more rows than columns.

(a) For arbitrary transformation matrix $A$, assume that $x$ is transformed into $y$, and that the $m_{ik}$ have been transformed into $m_{ik}'$. Find the new distance $d(y,m_{ik}')$ in terms of $A$, $y$, and $m_{ik}'$ and give an equation that relates $m_{ik}$ to $A$ and $m_{ik}$.
(b) What kind of distance is $d(y,m_{ik}')$ ? (Euclidean, Mahalanobis, or city block)
(c) Repeat part (b) if $A$ is an orthogonal matrix, where $A^{-1} = A^T$. 

Exam 2, EE5357, Summer 2006
4. In a nearest neighbor classifier, let $N_1$, denote the number of operations (additions, subtractions and multiplies) required to calculate one distance. The number of operations required to calculate all $K$ distances is then $N_o = K \cdot N_1$. Now, the partial distance $D_r(x, m_{ik})$ is defined as

$$D_r(x, m_{ik}) = \sum_{n=1}^{r} [x(n) - m_{ik}(n)]^2$$

For each $m_{ik}$, we find $D_r(x, m_{ik})$, and quit that center vector if $D_r(x, m_{ik}) > D_{\text{min}}$, the minimum distance found so far. Otherwise we must calculate the remaining part of the distance $d(x, m_{ik})$. A good assumption is that the number of center vectors where quitting occurs is $(r/N) \cdot K$

(a) Give an approximate expression for the number of required operations to calculate a partial distance, in terms of the symbols $r$, $N$, and $N_1$.

(b) For one input vector $x$, how many operations, on the average, are required to calculate the partial distances, for all center vectors where quitting occurs, in terms of $r$, $K$, $N$, and $N_1$.

(c) For how many center vectors does quitting not occur?

(d) How many operations are required to calculate the distances, for all center vectors where quitting does not occur?

(e) Give the total number of operations $N'_o$ for calculating the distances (part (b) plus part (d)), in terms of $K$, $N$, $r$, and $N_1$.

(f) Equating $d N'_o/dr$ to zero, find the value of $r$ that minimizes $N'_o$, and give the minimum $N'_o$ in terms of $N_1$ and $K$.

5. Given some Gaussian training data for a classification problem, a Bayes-Gaussian classifier and a polynomial discriminant (PD) classifier $d_i(x)$ can be designed through regression.

(a) Express the type B1 Bayes discriminant $u_i(x)$ in terms of the symbols $f(x \mid i)$ and $P_i$.

(b) Express $f(x)$, the joint pdf of $x$, in terms of the symbol $u_i(x)$.

(c) Express the type B2 Bayes discriminant $b_i(x)$ in terms of the symbol $u_i(x)$.

(d) Write $b_i(x)$ in terms of $m_i$, $C_i$, and its inverse $A_i$.

(e) Express the type B3 Bayes discriminant $d_i(x)$ in terms of the symbol $u_i(x)$.

(f) If $f(x \mid i)$ and $P_i$ are known, and $f(x \mid i)$ is Gaussian, which discriminant, $b_i(x)$ or the PD approximation to $d_i(x)$ (for $D=2$) is best?

6. In this problem, we have two classes ($N_c = 2$) and one Gaussian feature, $x$ ($N=1$). The class probabilities are not the same. $P_1 = .2$ and $P_2 = .8$.

(a) If $x$ has a different mean for each class ($m_1$ and $m_2$) and the same variance $\sigma^2$ for each class, give the Bayes-Gaussian discriminants. Simplify the discriminants if possible, and state their degrees.

(b) If $x$ has the same mean $m$ for each class and different variances ($\sigma_1^2$ and $\sigma_2^2$) for each class, give the Bayes-Gaussian discriminants. Simplify the discriminants if possible, and state their degrees.