Statistics 315a
Homework 2, due Wednesday February 8, 2012.

1. ESL 3.12 & 3.30
2. ESL 3.23
3. ESL 4.2
4. Suppose a_1 and a_2 are the first two discriminant coordinates, as outlined in (4.16) and below on page 116 in ESL. Let $Z_1 = a_1^T X$ and $Z_2 = a_2^T X$ be the corresponding discriminant variables. Show that Z_1 and Z_2 are uncorrelated (both the population and the sample versions).
5. Suppose you wish to build a regression model for some microarray data. In particular, you have 50 samples with expression measurements on each of 10,000 genes in a $50 \times 10,000$ array X. You also have an associated outcome variable y. You plan to use ridge regression to solve the problem.
   (a) Write an expression for the ridge regression coefficient vector $\hat{\beta}_\lambda$. How big is the matrix that requires inversion? What order of computations are required to compute $\hat{\beta}_\lambda$?
   (b) Show that $\hat{\beta}_\lambda$ is in the row-space of X i.e. $\hat{\beta}_\lambda = X^T\alpha$ for some α. [hint: start with the derivative condition on the penalized RSS criterion].
   (c) If $X = UDV^T$ is the SVD of X, argue that you can represent $\hat{\beta}_\lambda = V\theta$. Show how you can use this representation to dramatically reduce the computations in a), and for all values of λ.
   (d) How would you predict the outcome for a new sample with measurement vector x_0?
6. The zipcode data are high dimensional, and hence even linear discriminant analysis suffers from high variance. Using the training and test data for the 3s, 5s and 8s, compare the following procedures:
   (a) LDA on the original 256 dimensional space.
   (b) LDA on the leading 49 principal components of the features.
   (c) LDA when you *filter* the data as follows. Each non-overlapping 2×2 pixel block is replaced by its average.

(d) Multiple linear logistic regression using the same filtered data as in the previous question. [Use the `multinomial` family in the R package `glmnet`; use the solution at the end of the path].

Compare the procedures w.r.t training and test misclassification error.

Show analytically that the last of these procedures is equivalent to fitting a linear model with the full set of pixels, but the coefficients are constrained to be piecewise constant in 2×2 blocks [hint: represent the constrained coefficients as a linear expansion of *Haar* (piecewise constant) basis functions over the image domain.]

These data are available from the "Data" section of the book webpage: `http://www-stat.stanford.edu/ElemStatLearn`.

The last procedure produces a path of solutions, as we vary the penalty parameter λ for the elastic-net penalty. Produce a test-error trace as a function of the % deviance explained on the training data. Do this for several values of α. Summarize what you observe.