1. In finding the LDA transformation, the solution constrains $u^t W u = 1$ (similar to $u^t u = 1$ for PCA) and uses the lagrangian to get: $B u = \lambda W u$. For reasonable problems, $B$ is symmetric positive definite, and can be factored into $B = B^{1/2} B^{1/2}$. This allows us to solve a regular eigenvector problem for $v$:

$$B^{1/2} W^{-1} B^{1/2} v = \lambda v$$

and then find $u_k = B^{-1/2} v_k$ for the $k$-the eigenvector. When the true class-conditional distributions are Gaussian with shared covariance $\Sigma$, then the within class scatter is a scaled version of $W$. Since you can scale both $W$ and $B$, and not change the solution, we can use $\Sigma$ for $W$.

(a) Under the Gaussian shared-covariance assumption, find the new distributions $p(x' | \omega_i)$ of the transformed vector $x' = Ax$ where $A$ is a $p \times d$ LDA transformation, and the original class-conditional distributions are $N(\mu_i, \Sigma)$. You should find that the new vector elements are uncorrelated.

(b) If the covariances are not in fact the same, then $W$ is a weighted average of the different covariances. Assuming that the true original distributions are still Gaussian, will the transformed distribution be Gaussian? Will the elements be uncorrelated?

2. A classifier is said to have accuracy higher than 85% with 97.5% confidence for an error counting estimate. There were 9 errors on the test set. Roughly how many samples was it evaluated on? (Use the figure from your confidence interval class assignment, or equivalently DHS Fig 9.10. Note that the lower bound for a 95% two-sided confident region is the same as the lower bound for a 97.5% one-sided confidence region.) Another manufacturer reports that their classifier has accuracy of at least 89%, and they claim to have used the same test set. What might they have done differently to get a tighter tolerance for the same estimated error rate?

3. Using the file of scores and reference labels posted with the homework assignment, generate an ROC curve and a precision-recall curve to show performance, sweeping through the scores to test different thresholds.

4. You are designing a classifier that needs to be low cost, because it will be implemented in a cell phone and you don’t want to hurt the battery life too much. You have a fairly large amount of labeled training data. What types of approaches would you try and how would you choose between them?

5. What’s the VC dimension of the model set $M$ in a 2D space, where the model set $M$ includes all rectangles (and only rectangles), so that the decision regions induced by a model from set $M$ is one class inside a rectangular region, and one class outside the rectangular region?