Solution to the Midterm

1. (a) \[ \text{diag}(2, 1, 1) \]

(b) Call these vectors \( u \) and \( v \). Then, an example is:

\[
2uu' + 1/2 vv' = \begin{pmatrix} 5/4 & 3/4 \\ 3/4 & 5/4 \end{pmatrix}
\]

(c) Look at the picture below

![Image of two bell curves with blue and red colors]

Say blue is class 1 (with mean \( \mu_1 \)), red class 2 (with mean \( \mu_2 \)), with \( \mu_1 < \mu_2 \). The Bayes rule classifies to the closest \( \mu \). But an observation from population 1 has probability

\[
1 - \Phi((\mu_2 - \mu_1)/2)
\]

to be closer to \( \mu_2 \). Same thing for an observation from class 2 (reversing the roles of classes 1 and 2). As they have same proportion, the Bayes Risk is

\[
1 - \Phi((\mu_2 - \mu_1)/2) = \Phi(-|\mu_2 - \mu_1|/2)
\]

If we want that to be equal to 10%, we need

\[
|\mu_2 - \mu_1| = -2\Phi^{-1}(0.10) = 2.5631
\]
2. (a) **True** Take a $3 \times 3$ covariance matrix with rank 2. Choose an orthobasis \( \{a_1, a_2, a_3\} \) corresponding to its eigenvalues in decreasing order (the last eigenvalue is 0). Then the data lives in \( \text{span}(a_1, a_2) \) and therefore \( a_i^i x_i \) is constant across \( i \).

(b) **False** If your data is standardized, those matrices are the same.

(c) **True** In fact \( a_2 \) is not even defined uniquely.

(d) **False** In general the proportions will be different, and the Bayes classifier takes that into account.

(e) **False** It does not obey the triangle inequality. The square root of the Mahalanobis distance does.

(f) **False** But true if \( \Sigma^{-1} \) is replaced by \( \Sigma^{-1/2} \).

The last two were controversial, and not taken into account in the grading process.

3. (a) \$0.1, since guessing that everything comes from the larger class would result in 10% error, hence \$0.1 expected loss. The Bayes risk cannot be larger, but we can certainly think up examples where it is exactly \$0.1.

(b) \((\mu_1 - \mu_2)\Sigma^{-1}(x - 1/2(\mu_1 + \mu_2)) > 0\)

(c) Assume WLOG that \( \mu = 0 \). The two likelihood functions:

\[
\begin{align*}
    f_1(x) &= (2\pi)^{-\frac{3}{2}} |\Sigma_1|^{-\frac{1}{2}} \exp(-\frac{1}{2} x'\Sigma_1^{-1} x) \\
    f_2(x) &= (2\pi)^{-\frac{3}{2}} |\Sigma_2|^{-\frac{1}{2}} \exp(-\frac{1}{2} x'\Sigma_2^{-1} x)
\end{align*}
\]

Now, \(|\Sigma_2|^{-\frac{1}{2}} = 2^{-\frac{3}{2}} |\Sigma_1|^{-\frac{1}{2}}\), and \(\exp(-\frac{1}{4} x'\Sigma_2^{-1} x) = \exp(-\frac{1}{4} x'\Sigma_1^{-1} x)\).

Hence the desired condition to assign to class 2 is:

\[
\frac{2^{-\frac{3}{2}} \exp(-\frac{1}{2} x'\Sigma_1^{-1} x)}{\exp(-\frac{1}{2} x'\Sigma_1^{-1} x)} > 1
\]

\[
\frac{1}{4} x'\Sigma_1^{-1} x > \frac{p}{2} \log 2
\]

\[
x'\Sigma_1^{-1} x > 2p \log 2
\]

4. (a) There are of course countless examples of data on which Discriminant Analysis could be run. Consider for example identifying email spam, where the input data is a large number (thousands or even millions) of “good” and “bad” emails, the features (predictors) may be words or phrases, email addresses etc.. For example, the words “sex” or “prize” may be good predictors indicating spam, while the email extension “.edu” may indicate a legitimate email. We can expect to have thousands of features. If our input features are counts of the number of appearances of each word or phrase we cannot expect them to be distributed normally. A poisson distribution may be more likely.
Maximum likelihood classifiers require assumptions about the distribution of the input data. As we mentioned, the data cannot reasonably be expected to be normal, and hence likelihood solutions assuming normal distribution would probably be inappropriate. We could devise multi-Poisson models, which we may be able to estimate under simplifying assumptions.

Bayesian approaches may be a good idea in this case, since the large number of predictors would require some kind of variable or model selection, which is easier to do in a Bayesian framework. Of course, to use this approach we would have to devise appropriate priors for the multi-variate distribution of our predictor vectors in the two classes. Fisher’s LDA can be a reasonable approach (if we ignore its strong connection to the normal likelihood rule and treat it as a non-parametric approach). The problem with it may be the curse of dimensionality - the need to calculate $\Sigma^{-1}$ in high dimension and the unreliability of the result.

Nearest neighbor approaches may be a good solution to this problem. Applying them successfully would require a “good” metric for the distance between observations. We could try to devise such a metric based on experts’ understanding of the problem and the data.