FINAL EXAM: Problem 2

STAT 472, Spring 2020

The following will read in the appropriate data:

```
train.dat = read.table("http://mason.gmu.edu/~csutton/train2.txt", header=TRUE)
dim(train.dat)
head(train.dat)
plot(train.dat$x,train.dat$y,col=train.dat$class+2)
```

The scatter plot has 84 class 0 observations plotted in red, and 112 class 1 observations plotted in green.

Due to what appears to be differences in the variances and perhaps the correlations, it may be that QDA will be superior to LDA. But if the differences are small enough, then it may be that LDA will perform better, since with it there are fewer parameters to estimate. Let's fit an LDA model and then use the training data to get what is sometimes called a "resubstitution estimate" of the error rate (which may be overoptimistic).

```
library(MASS)
lda.fit=lda(class~x+y,data=train.dat)
lda.pred=predict(lda.fit, train.dat)
table(lda.pred$class, train.dat$class)
mean(lda.pred$class != train.dat$class)
```

The estimated error rate is about 0.036.

Now use the available data to fit a QDA model.

(a) (1 point) What is the "resubstitution estimate" of the error rate, rounded to the nearesr thousandth?

You should find that the estimate from the fitted QDA model is a little larger than the corresponding estimate from the LDA fit. But since both models may overfit the data, it may be better to rely on estimates obtained from cross-validation, or from using the validation set approach described in subsection 5.1.1 of *ISL*.

To use the validation set approach, we can randomly select 65 observations (about 1/3 of the available data) to serve as a validation set, and then see how well models fit on the remaining 131 observations predict the class for the members of this validation set.

```
RNGversion("3.6.2")
set.seed(123)
val = sample(196,65,replace=FALSE)
lda.fit=lda(class~x+y,data=train.dat[-val,])
lda.pred=predict(lda.fit, train.dat[val,])
table(lda.pred$class, train.dat$class[val])
mean(lda.pred$class != train.dat$class[val])
```

The estimated error rate is about 0.046.

Now use the validation set chosen above to estimate the QDA error rate.

(b) (1 point) What is the estimated error rate, rounded to the nearesr thousandth?

The trouble with the validation set approach estimates is that the models are fit using only 131 observations, and also there are only 65 independent observations with which to get the predictions needed for the estimates. Because of these small numbers, it may be better to use cross-validation.

To use cross-validation to estimate the error rates, we should first randomize the order of the 196 observations.

```
set.seed(135)
random.order = sample(196,196,replace=FALSE)
new.train=train.dat[random.order,]
```

Now use 7-fold cross-validation to obtain an estimate of the QDA error rate. Use the first 28 observations in `new.train` for the first fold, the second 28 observations in `new.train` for the second fold, and so on.

(c) (2 points) What is the estimated error rate, rounded to the nearesr thousandth? (If you use cross-validation to estimate the LDA error rate, you should get a value of about 0.036. As another check, I'll give you that the QDA error rate estimated from a rather large independent test sample is about 0.060, and the cross-validation estimate of the error rate is closer to that value than are the estimates obtained by resubstitution and the validation set approach (using a validation set of only 65 randomly chosen observations).)