Math 583 HW 8 Fall 2017. Due Wednesday, Nov. 1.
Quiz 8 on Friday, Nov. 3 is similar to HW 8. Use 3 sheets of notes.

Problem numbers and example numbers are from the Olive text.

```
Logistic Regression Output for Problem 5.2
Response = nodal involvement, Terms = (acid size xray)
Label      Estimate  Std. Error     Est/SE    p-value
Constant  -3.57564   1.18002        -3.030     0.0024
acid       2.06294   1.26441         1.632     0.1028
size       1.75556   0.738348        2.378     0.0174
xray       2.06178   0.777103        2.653     0.0080

Number of cases: 53, Degrees of freedom: 49,
Deviance: 50.660
```

**A) 5.2.** Following Collett (1999, p. 11), treatment for prostate cancer depends on whether the cancer has spread to the surrounding lymph nodes. Let the response variable = group $y$ = *nodal involvement* (0 for absence, 1 for presence). Let $x_1 = acid$ (serum acid phosphatase level), $x_2 = size$ (= tumor size: 0 for small, 1 for large) and $x_3 = xray$ (xray result: 0 for negative, 1 for positive). Assume the case to be classified has $\boldsymbol{x}$ with $x_1 = acid = 0.65$, $x_2 = 0$, and $x3 = 0$. Refer to the above output.

a) Find ESP for $\boldsymbol{x}$.
b) Is $\boldsymbol{x}$ classified in group 0 or group 1?
c) Find $\hat{\rho}(\boldsymbol{x})$.

Hint: $ESP = \hat{\alpha} + \sum_{i=1} x_i \hat{\beta}_i$ where $\hat{\alpha}$ corresponds to the constant and the $x_i$ are nontrivial predictors.

**B) 5.3.** Recall that $X$ comes from a uniform(a,b) distribution, written $x \sim U(a, b)$, if the pdf of $x$ is $f(x) = \dfrac{1}{b-a}$ for $a < x < b$ and $f(x) = 0$, otherwise. Suppose group 1 has $X \sim U(-3, 3)$, group 2 has $X \sim U(-5, 5)$, and group 3 has $X \sim U(-1, 1)$. Find the maximum likelihood discriminant rule for classifying a new observation $x$.

Hint:See example done in class.

```
> out <- lda(x,group) #Problem 5.5
> 1-mean(predict(out,x)$class==group)
[1] 0.02
>
> out<-lda(x[,-c(1)],group)
> 1-mean(predict(out,x[,-c(1)])$class==group)
[1] 0.02
> out<-lda(x[,-c(1,2)],group)
> 1-mean(predict(out,x[,-c(1,2)])$class==group)
[1] 0.04
```

```
> out<-lda(x[,-c(1,3)],group)
> 1-mean(predict(out,x[,-c(1,3)])$class==group)
[1] 0.03333333
> out<-lda(x[,-c(1,4)],group)
> 1-mean(predict(out,x[,-c(1,4)])$class==group)
[1] 0.04666667
>
> out<-lda(x[,c(2,3,4)],group)
> 1-mean(predict(out,x[,c(2,3,4)])$class==group)
[1] 0.02
```

**C) 5.5.** The above output is for LDA on the famous iris data set. The variables are $x_1$ = sepal length, $x_2$ = sepal width, $x_3$ = petal length, and $x_4$ = petal width. These four predictors are in the $x$ data matrix. There are three groups corresponding to types of iris: setosa, versicolor, and virginica.

a) What is the AER using all 4 predictors?

b) Which variables, if any, can be deleted without increasing the AER in a)?

Hint: See Example 5.2.

**Do the two source commands for sldata and slpack.**

**D) 5.11.** The Wisseman et al. (1987) pottery data has 36 pottery shards of Roman earthware produced between second century B.C. and fourth century A.D. Often the pottery was stamped by the manufacturer. A chemical analysis was done for 20 chemicals (variables), and 28 cases were classified as Arrentine (group 1) or nonArrentine (group 2), while 8 cases were of questionable origin. So the training data has $n = 28$ and $p = 20$.

a) Copy and paste the $R$ commands for this part into $R$ to make the data set.

b) Because of the small sample size, LDA should be used instead of QDA. Nonetheless, variable selection using QDA will be done. Copy and paste the $R$ commands for this part into $R$. The first 9 variables result in no misclassification errors.

c) Now use commands like those shown in Example 5.2 to delete variables whose deletion does not result in a classification error. You should get four variables are needed for perfect classification. What are they (e.g. X1, X2, X3, and X4)?

(The classification error and perfect classification are for the training data, not the test data. So AER = 0. For c) only some of the code is given, you need to add some similar commands. See Example 5.2.)

**E) 5.12.** Variable selection for LDA used the pottery data described in Problem 5.11, and suggested that variables X6, X11, X14, and X18 are good. Use the $R$ commands for this problem to get the apparent error rate AER.

**F) 5.13.** This problem uses KNN on the same data set as in Problem 5.11.

a) Copy and paste the commands for this part into $R$ to show AER = 0 for KNN if $K = 1$.

b) Copy and paste the commands for this part into $R$ to get the validation error rate

for KNN if $K = 1$. Give the rate. The validation set has 12 cases and KNN is computed from the remaining 16 cases.

c) Use these commands to give the AER if $K = 2$.

d) Use these commands to give the validation ER if $K = 2$.

e) Use these commands to give the AER for 2NN using variables $X6, X11, X14$, and $X18$ that were good for LDA in Problem 5.11.

f) Use these commands to give the validation ER for 2NN using variables $X6, X11, X14$, and $X18$ that were good for LDA.