Homework 2 (4570, 4880)

Due Friday March 1 by 2 pm

★ Turn in a typed hardcopy of your explanations, relevant output, and R code by the due date and time.

Titanic Data. This data set (`titanic_ctrain`) contains information on the output variable for whether (1) or not (0) passengers survived (`Survived`) the fatal maiden voyage of the Titanic ocean liner. The training data was cleaned to include 534 passengers and the following 6 input variable variables. Please keep original variable coding, except for scaling of the input variables.

```
Pclass = passenger class (1st, 2nd, 3rd)
Sex = gender of passenger (0: female, 1: male)
Age = age in years of the passenger
SibSp = number of siblings plus spouse of the passenger
Parch = number of parents/children of the passenger
Fare = the price in British pounds for the passenger
```

H.2.1 Fit the first-order (FO) Logistic Regression (LR) model.

(a) Perform the hypothesis test associated with `Pclass` at $\alpha = 0.05$. Give the hypotheses, test statistic, statistical decision, and practical conclusion of the test.

(b) Interpret the odds estimate and associated 95% confidence interval for `Pclass`.

(c) Give the confusion matrix and error rate using *resubstitution*. Explain what the confusion matrix indicates about the types of prediction errors made by the logistic regression model.

H.2.2 Fit all input variables for LDA, QDA, and KNN. Use proportional priors for Discriminant Analysis and standardize the predictors for KNN.

(a) Which variables are most important for discrimination with LDA? Explain.

(b) Assess the assumption of equal population covariance matrices at $\alpha = 0.05$. Give the statistical hypotheses, test statistic, statistical decision, and practical conclusion of the test. Should QDA predict better than LDA given this test result and the discussion in Section 4.5 (Ch 4nts, p 6)?

(c) Compare the LOOCV error rates for LDA, QDA, and KNN (with K = 5). Which method appears to be better in predicting Survival?

R-Code

You can read the **Titanic Data** into R by downloading the file titantic.csv from WyoCourses-Files/DATA to a folder called DataMining and running the following code.

```
> dat = read.csv('C:/DataMining/titanic.csv')

> dim(dat)
[1] 534   7

> names(train)
[1] "Survived" "Pclass" "Sex" "Age" "SibSp" "Parch" "Fare"

> X = cbind(dat$Pclass,dat$Sex,dat$Age,dat$SibSp,dat$Parch,dat$Fare)
> standard.X = scale(X)
> y = dat$Survived
```