Assignment 3. ST464/ST684
Catherine Hurley
Due Friday April 2, 6pm.

- Do all questions.
- Upload solution to question 2 only to Moodle by Friday April 2nd, 6pm.(This is Good Friday. You are welcome to hand in early!.)
- You should complete the assignment in Rmarkdown, knit the file to html and upload the html file to Moodle.
- Please use the skeleton Rmarkdown file h3student.Rmd available on Moodle.
- Fill in your name and student number on the solution file in the space provided.
- The upload must be completed by time and date given above or the assignment will not be accepted.
- The tutorial on week 8 will provide assistance with this material.

1. Suppose we wish to predict whether a given stock will issue a dividend this year (yes or no) based on X, last year's percentage profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was 10, while the mean for those that didn't was 0. In addition, the variance of X for these two sets of companies was 36. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.
   Recall that the normal probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

   where μ is the mean and σ^2 is the variance.

   Hint: recall the formula

$$P(Y = j|X = x_0) = \frac{\pi_j f_j(x_0)}{C}$$

2. For the Pima diabetes data from package mlbench, split the data into a test and training sets of size containing 50% and 50% of observations each, using the code below.

```
 # install.packages("mlbench")  # first time only
library(mlbench)
 library(ggplot2)
 library(class)
 library(MASS)
data(PimaIndiansDiabetes2)

d <- na.omit(PimaIndiansDiabetes2)
set.seed(2)
s <- sample(nrow(d), round(.6*nrow(d)))
dtrain <- d[s,]
dtest<- d[-s,]
```

   (a) Plot the variables age and glucose using colour to show the two levels of diabetes for the training set.

[2 marks]

(b) Perform a logistic regression analysis to predict diabetes, using variables age and glucose, on the training set. Use a plot to show the logistic classification boundaries and the training data. What is the test error of the model obtained?

[3 marks]

(c) Perform a linear discriminant analysis to predict diabetes, using variables age and glucose, on the training set. Use a plot to show the discriminant boundaries and the training data. What is the test error of the model obtained?

[3 marks]

(d) Repeat (b) using quadratic discriminant analysis. Which is better, logistic, LDA or QDA?

[4 marks]

(e) Perform KNN with response of diabetes, and the same two predictors. Remember to scale the predictors for the training set, and apply this scaling to the test set. Use $k = 5$ and $k = 30$. Which value of k gives the best result on the test set?

[5 marks]

(f) For the better value of k plot the training data and the classification boundaries from knn. Which classification algorithm would you recommend here based on your findings?

[3 marks]

3. Use the diabetes data from the previous question split into training and testing subsets.

   (a) Perform a logistic regression to predict diabetes, using variables glucose, age, mass, insulin on the training set. What is the test error of the model obtained?

   (b) Redo (a) using linear discriminant analysis.

   (c) Repeat (b) using quadratic discriminant analysis.

   (d) Perform KNN for this problem. Remember to scale the predictors for the training set, and apply this scaling to the test set. Use $k = 5$ and $k = 30$. Which value of k gives the best result on the test set? Which method of logistic, lda, qda, knn gives the best result?

4. A classifier gives the following result. In the table below, Group gives the true class, and Prob gives the estimated probability of Group A (positive) using the classifier.

	Group	Prob
1	A	0.486
2	A	0.560
3	A	0.701
4	A	0.936
5	A	0.441
6	A	0.593
7	B	0.623
8	B	0.436
9	B	0.415
10	B	0.041

(You can do this question in R or by hand)

   (a) What are the predicted classes? Use a threshold of 0.5.

   (b) What is the error rate? What is the false positive rate? The true positive rate?

   (c) Now let the threshold take values 0, .2, .4,.6,.8,1. For each threshold calculate the false positive rate, and the true positive rate. (If doing this in R use more thresholds.)

(d) Plot the true positive rate versus the false positive rate. This is the ROC curve.

(e) (Optional, if doing in R) Another classifier just assigns class probabilities randomly, ie the estimated probabilities are:

	Group	Prob
1	A	0.206
2	A	0.177
3	A	0.687
4	A	0.384
5	A	0.770
6	A	0.498
7	B	0.718
8	B	0.992
9	B	0.380
10	B	0.777

Plot the ROC curve for this classifier.