Stat 627/427 (Statistical Machine Learning)

Homework # 5
Classification methods: Logistic regression, KNN, LDA, QDA (sec. 2.2.3, 3.5, ch. 4)

1. **Pages 53–54, chap. 2, #7.** The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

2. **Pages 169–170, chap. 4, #5.** We now examine the differences between LDA and QDA.

(a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

(b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

(c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

(d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

3. **Page 170, chap. 4, #7.** Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. 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.
Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$. *You will need to use Bayes' theorem.*

4. **Page 171, chap. 4, #10(b-d, e-h + additional i, j).** Here we try to predict behavior of the market next week. The *Weekly* data set contains 1089 observations with the following 9 variables.

Year	The year that the observation was recorded
Lag1	Percentage return for previous week
Lag2	Percentage return for 2 weeks previous
Lag3	Percentage return for 3 weeks previous
Lag4	Percentage return for 4 weeks previous
Lag5	Percentage return for 5 weeks previous
Volume	Volume of shares traded (average number of daily shares traded in billions)
Today	Percentage return for this week
Direction	A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

This data set is a part of ISLR package, and you can obtain it by typing

```
> install.packages("ISLR")
> library(ISLR)
> attach(Weekly)
```

(b) Perform a logistic regression with *Direction* as the response and the five lag variables plus Volume as predictors. Do any of the predictors appear to be statistically significant? If so, which ones?

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

(d) *Cross-validation...* Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(i) Plot an ROC curve for the logistic regression classifier, using different probability thresholds.

(j) Will KNN method perform better? Use all five Lag variables and predict the direction of the market in 2009-2010 based on training data 1990-2008. Try different k and select the optimal one. Give a confusion matrix.

(e) Use LDA with a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

(f) Repeat (e) using QDA.

(g) Repeat (e) using KNN with $K = 1$.

(h) Which of our classification methods appears to provide the best results on this data?