Statistical Data Analysis
Exercise Session 9

Exercise 1: Logistic regression.
Consider the dataset titanic.txt about passengers on the Titanic's ill-fated voyage. The variables are:

variable	interpretation
Surv	whether the passenger survived
Pclass	passenger class
Gender	passenger's gender
Age	passenger's age (in years)
SibSp	number of siblings+spouse on board
Parch	number of parents+children on board
Fare	fare paid for the voyage

1. Fit a logistic regression model that uses the regressor variables to predict the probability of survival. Interpret the results.
2. Apply stepwise variable selection to get a more parsimonious model. Use the AIC criterion. Note that the AIC is still defined as $AIC_p = -2 \max \log L + 2p$ as in Chapter 6, hence it becomes Residual deviance + $2p$ in logistic regression.
3. Can you conclude that the more parsimonious model can be retained based on a deviance test?
4. Use the reduced model to compute the odds ratio and its 95% confidence interval for the effect of a 1 year increase in Age.
5. Plot both the observed survival and its predicted probability as a function of $\hat{\eta}_i = \mathbf{x}_i \hat{\beta}$
6. Make a plot of the deviance residuals. Are there outlying residuals?
7. Suppose you set the predicted outcome to 1 if and only if $\hat{\pi}_0 > c$ where c is the fraction of 1's among the observed survival in the data. What is then the apparent error rate APER of this model?

Exercise 2: LDA and QDA.
Consider the abalone training dataset (abalone1.txt) where:

variable	type	unit	interpretation
Gender	nominal	M, F	
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Weight	continuous	grams	weight of whole abalone
Shucked	continuous	grams	weight of meat
Viscera	continuous	grams	gut weight
Shell	continuous	grams	after being dried

1. Use LDA to fit a linear discrimination rule for predicting Gender.
2. Use partimat() to make a matrix plot of the linear discriminant func- tion.
3. Compute the APER, the LOO misclassification error and the misclassification error of the test data (abalone2.txt) to evaluate the misclassification rate of the discriminant rule. Compare the different values.

4. Repeat the previous 3 questions but this time using QDA instead of LDA. Compare the results.

Exercise 3 : k-Nearest Neighbor rule.
Consider again the abalone dataset (abalone1.txt).

1. Carry out k-NN in which the test set is taken equal to the training set, for k = 1 and k = n, and compute the confusion matrices. Explain the results.
2. Use L00-CV to select an optimal value of k.
3. Compare the LOO performance of the k-NN rule with the actual classification performance on the test data (abalone2.txt).
4. All the algorithms seem to perform equally well. What might be the reason?

Some useful functions:

1. glm() en prop.table() in the stats library.
2. lda() en qda(), predict.lda() and predict.qda() in the MASS pack- age.
3. partimat() in the klaR package.
4. knn and knn.cv in the class package.