CV Homework 4

Code

This is an R script with the purpose of running logistic regression and naive bayes on the AppliedPredictiveModeling abalone dataset

Ramesh Kanakala

Step 1

Hide

```
#load mlbench and take a look at ablo
library(AppliedPredictiveModeling)
data(abalone)
ablo <- abalone[abalone$Type != "I", ]
summary(ablo$Type)
str(ablo)
head(ablo)

#percent of each Type
typesum <- summary(ablo$Type)
male <- typesum[[1]]/sum(typesum)
fem <- typesum[[3]]/sum(typesum)
print(paste("Male % = ", male))
print(paste("Female % = ", fem))
```

a. There are 2835 instances in the ablo data set (removed rows of Type 'I' to keep it binary Typeification)
b. The target column is Type: whether the instance is Male (1) or Female (3)
c. There are 8 predictors, all with a numerical data type
d. 46.1% of the observations are male and 53.9% are female

Step 2

Hide

```
#build model
glm0 <- glm(Type~Diameter+LongestShell, family = "binomial", data = ablo)
summary(glm0)
```

Didn't receive the error (glm.fit: fitted probabilities numerically 0 or 1 occurred) as the independent variables may not be differentiating the dependent perfectly well.

Step 3

Hide

```
#adding Diameter.Big and LongShell.Big columns
ablo$Diameter.Big <- as.factor(ifelse(ablo$Diameter>0.4600, 1, 0))
ablo$LongShell.Big <- as.factor(ifelse(ablo$LongestShell>0.5850, 1, 0))
summary(ablo$Diameter)
summary(ablo$LongestShell)
summary(ablo$Diameter.Big)
summary(ablo$LongShell.Big)
```

The new columns seem to be evenly distributed, or balanced. I believe creating these new binary columns were a good idea as a balanced data set usually means there are less problems for classication algorithms and greater accuracy.

Step 4

Hide

```
#conditional density plots with Diameter and LongestShell
attach(ablo)
par(mfrow=c(1,2))
cdplot(y = Type, x = Diameter)
cdplot(y = Type, x = LongestShell)
```

We can see for both diameter and type, as well as longest shell measurement and type, the areas are very similar in size and shape. I think the cutoff point I chose, each variable's respective median, for both diameter and longestshell was justified as we can see the conditional density areas vary differently around 0.5.

Step 5

Hide

```
#plots with new columns
par(mfrow=c(1,2))
plot(y = Type, x = Diameter.Big, xlab = "Diameter Big Factor", ylab = "Gender Type")
plot(y = Type, x = LongShell.Big, xlab = "LongShell Big Factor", ylab = "Gender Type")
cdplot(y = Type, x = Diameter.Big)
cdplot(y = Type, x = LongShell.Big)
sum(Type=='M' & Diameter.Big==1)/nrow(ablo)
sum(Type=='M' & Diameter.Big==0)/nrow(ablo)
sum(Type=='M' & LongShell.Big==1)/nrow(ablo)
sum(Type=='M' & LongShell.Big==0)/nrow(ablo)
```

The conditional plots don't really exemplify the stark difference of gender from a small diameter or not abalone and a small shell measurement or not abalone, however, both plots and cdplots do show a slight difference between the two with small diameters having more males. The areas for both graph are very close. We see that males are more associated with small diameters and longest shell measurements both with the plots as well as the percentages. There is not a great difference of the small diameters and longest shell measurements being male, however.

Step 6

Hide

```
#divide ablo into train and test sets
set.seed(1234)
i <- sample(1:nrow(ablo), nrow(ChickWeight)*0.8,
replace=FALSE)
train <- ablo[i,]
test <- ablo[-i,]
```

Step 7

Hide

```
#logistic regression classifier for Type given Diameter.Big and LongShell.Big
glm1 <- glm(Type~Diameter.Big+LongShell.Big, data=train, family=binomial)
summary(glm1)
```

a. Both Cell small and regular seem to be bad predictors as the have quite high p-values.
b. The deviance dropped slightly from null to residual meaning that the predictors did make a difference in the model.
c. An AIC by itself is not very useful but comparing it to the previous model, it decreased greatly meaning that this model is better.

Step 8

Hide

```
#test the model on the test data and compute accuracy
library(e1071)
probs <- predict(glm1, newdata=test, type="response")
pred1 <- ifelse(probs>0.5, 'M', 'F')
acc <- mean(pred1==test$Type)
acc
library(caret)
confusionMatrix(as.factor(pred1), test$Type)
```

The model has an accuracy of 0.5179; not very good. There were essentially the same amount of false negatives than false positives. Perhaps diameter and longest shell measurements aren't the best predictors for gender.

Step 9

Hide

```
#coefficients
glm1$coefficients[]
smallodds <- exp(glm1$coefficients[2])
smallprob <- (smallodds)/(1 + smallodds)
smallprob
sum(Type=='M' & Diameter.Big==1)/nrow(ablo)
```

a. The coefficient for cell small is -0.5319929 and for cell regular -0.1566952
b. The coefficents are in log odds which means that for a unit increase in in Diameter and LongShell, the Type decreases by log odds of -0.5319929 and -0.1566952

c. The estimated probability of malignant if Cell.small is true is 0.3700522
d. The probability of male if Diameter.Big is true is 0.2440917. It's somewhat close to the estimated probability but a little less; the estimation works with a smaller part of the data, the train set, but the calculated probability is with the entire set of data which may be why it is slightly different but somewhat close.

Step 10

Hide

```
#two more models using just Diameter.Big or LongShell.Big
glm_diameter <- glm(Type~Diameter.Big, data=train, family=binomial)
glm_longshell <- glm(Type~LongShell.Big, data=train, family=binomial)
anova(glm_diameter, glm_longshell, glm1)
summary(glm_diameter)
summary(glm_longshell)
```

Taking a look at the anova, we see that that model 3 has the lowest residual deviance (by just 1, however) suggesting that combining both variables improves the model. Looking at the AIC scores from the model summaries for each model, we see that Diameter.Big has the lowest here as well.

Step 11

Hide

```
#naive bayes model with Diameter.Big and LongShell.Big
library(e1071)
nb1 <- naiveBayes(Type~Diameter.Big+LongShell.Big , data=train)
nb1
summary(nb1)
```

a. 48.7% of the training data is female
b. The likelihood a male sample is not small is 0.5738397
c. The likelihood a male sample is not regular is 0.5738397

Step 12

Hide

```
#predict with naive bayes model
pred2 <- predict(nb1, newdata=test, type="class")
acc <- mean(pred2==test$Type)
acc
confusionMatrix(pred2, test$Type)
```

The accuracy and confusion matrix output is essentially the same as the logistic regression model with the naive bayes model. This may be because the Type frequencies of malignant and benign are somewhat unbalanced with almost 64% being malignant leading to the classifiers reaching close accuracies. Essentially, the models classify very similarly based on the two shared predictors.