Homework 4 Possible Solutions

Jessica Kraker

Possible Solutions to Selected Questions

Important: This document contains R code solutions and example answers for problems posed in the DS740: Data Mining homework. We are intentionally sharing this document with learners *who have already completed the associated homework*. We want you to be able to review and troubleshoot your code, as well as ask questions to prepare you for later assessments. *By using this document, you are accepting responsibility* **not** *to share it with anyone else, including other students in the course or the program who might not have completed the homework yet*. By upholding this agreement, you are helping us use this tool with learners in future terms.

From Problem 1: Use LDA with One Predictor

Question

Define a new variable called Domestic to have the value 1 when the car is domestic (*origin* = 1, for American) and the value 0 when the car is foreign (*origin* = 2 or 3, for European or Japanese, respectively). Tabulate the results, and report the count of domestic (*Domestic*=1) vehicles.

Possible Code Answer

```
NewAuto <- Auto %>%
  mutate(Domestic = as.numeric(origin == 1))  # given
# possible answer
table(NewAuto$Domestic)
```

```
##
##   0   1
## 147 245
```

Question

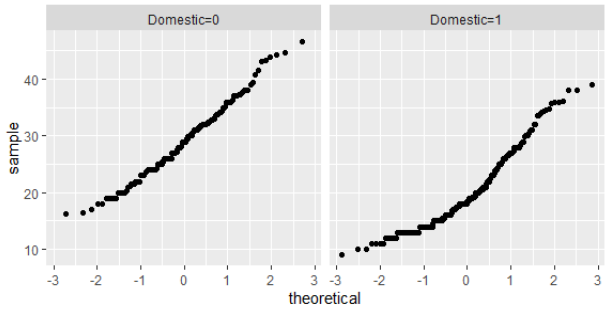
Compute and enter the mean *mpg* for domestic and foreign vehicles

Possible Code Answer:

```
# Question 8
NewAuto %>%
  group_by(Domestic) %>%
  summarize(MeanMPG = mean(mpg))
```

Question

Normal probability plots of mpg for the two groups are shown in the image, for each level of Domestic. Using these, along with the boxplot you produced earlier, discuss whether the two assumptions for running an LDA are reasonably met.



Possible Answer: *The spreads of the two sets of mpg seem very similar (note the standard deviations are 6.46 and 6.44 respectively for Foreign and Domestic vehicles).*

As for normality, neither boxplot shows strong outliers. And the normal probability plot for Foreign vehicles shows a straight line, suggesting normality is reasonable [Anderson-Darling test for non-normality is insignificant, with P-value of 0.2718]. However, the normal probability plot for Domestic vehicles shows a curve, suggesting normality is NOT plausible due to skewness of the data [Anderson-Darling test for non-normality is highly significant, with P-value of 1.46e-12].

Comment: We will proceed with LDA, but note that careful, honest assessment must be done to validly determine the utility of the model.

Question

Fit the linear discriminant analysis, using the predictor *mpg* to predict the response *Domestic*.

Predict the classifications from the LDA fit and tabulate the variable *Domestic* with the LDA classification; compute sensitivity and specificity

Possible Code Answer:

```
# Question 6
ldafit1 = lda(Domestic~mpg,data=NewAuto)

fittedclass = predict(ldafit1,data=NewAuto)$class
table(NewAuto$Domestic,fittedclass)

library(caret)

confusionMatrix(fittedclass,as.factor(Domestic),positive="1")
```

Question

Would you prefer to use LDA or QDA when using the variable selected to predict Domestic? Explain your reasoning, with particular reference to differences in assumptions between the two methods.

Possible Answer: *QDA - there is clearly a strong difference in the variabilities of the Domestic versus Foreign values.*

Question

Based on your answers to the previous two questions, fit the discriminant analysis method you selected, using your selected variable to predict the response *Domestic*.

Produce a ROC curve for this fitted model. Use the "Embed Image" button to upload your plot in the Canvas homework question.

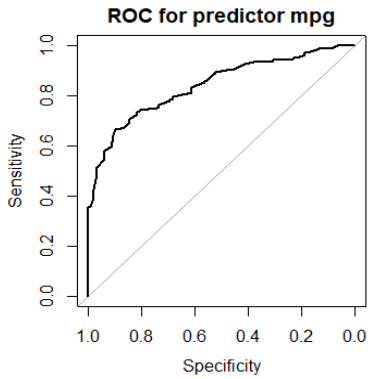
Possible Code Answer:

```
qdaprob = predict(qda(Domestic~displacement,
data=NewAuto),data=NewAuto)$posterior[,2]
qda.roc <- roc(response=NewAuto$Domestic, predictor=qdaprob)

plot.roc(qda.roc); auc(qda.roc)
```

Question

The ROC curve and R output for using *mpg* to predict *Domestic* is shown in the image. Discuss which variable (*mpg* or the variable you selected previously) that you would use to predict *Domestic* and **why**.



Possible Answer: I would use the variable I selected, since it appears to be more distinguishing (larger area under curve). Note that using QDA is needed due to differences in variances, but actually has lower AUC than with LDA; this may be due to the normality assumption not being met. Thus, again, careful assessment is required.

From Problem 2: Use LDA & QDA with Multiple Predictors

Question

Make the variable *origin* into a factor. Then, produce a scatterplot of *mpg* and *displacement*, marked by *origin*, along with an appropriate legend. Use the "Embed Image" button to upload your plot in the Canvas homework question.

Possible Code Answer:

```
NewAuto <- NewAuto %>% mutate(origin = factor(origin))
ggplot(NewAuto, aes(x = mpg, y = displacement)) +
    geom_point(aes(color = factor(origin)))
```

Questions

Fit the linear discriminant analysis, using the predictors *mpg*, *cylinders*, *displacement*, *horsepower*, and *weight* to predict the response *origin.* Predict the classifications from the **LDA** fit. Cross-tabulate the variable *origin* with the LDA classification, and report the number of **correctly** classified vehicles, for each of American, European, and Japanese. You may find `help(Auto)` useful for the meaning of levels of *origin.*

Possible Code Answer:

```
ldafit2 = lda(origin~mpg+cylinders+displacement+horsepower+weight, data=NewAuto)
fittedclasslda2 = predict(ldafit2,data=NewAuto)$class
table(NewAuto$origin,fittedclasslda2)
```

Question

Describe how the predictive abilities (as assessed on the original data) compare between LDA and QDA fits. Discuss why these results seem reasonable, given plots of various predictors marked by *origin* (such as the one you made at the beginning of this problem).

Possible Answer: *The classification rates are comparable for American and European vehicles, but the QDA model is clearly more accurate (87.3%) than LDA (72.2%) for Japanese vehicles. This makes sense, given the differences in variability of predictors, grouped by origin. Different covariance matrices suggest using QDA over LDA.*

Question

Using the **QDA** fit, for a vehicle which has 20 *mpg*, 8 *cylinders*, *displacement* of 320 in^3, 280 *horsepower*, and *weight* of 3600 pounds, predict the *origin* of the vehicle.

Possible Code Answer:

```
newdata = data.frame(mpg=20, cylinders=8, displacement=320, horsepower=280,
weight=3600)
predict(qdafit2,newdata)$class
```

From Problem 3: Model Selection with LDA and QDA

Question

Use the below code to set R's seed to 4 and define **cvgroups** (random groups for the cross-validation) using the `sample()` function.

With **cvgroups** as just defined, use 10-fold cross-validation method to calculate $CV_{(10)}$ for each of Models 1-4. Include all your code (that is, the full loop process) for computing honest predictions and the $CV_{(10)}$ measure for the four models.

Possible Code Answer:

```
nfolds = 10; n=392
groups = rep(1:nfolds,length=n)
set.seed(4)
cvgroups = sample(groups,n); table(cvgroups)

# code for question 22
CVmodels1_4 = rep(NA,4)
ModelA = origin~displacement
ModelB = origin~mpg+cylinders+displacement+horsepower+weight
MethodsUsed = c("LDA","LDA","QDA","QDA")
ModelsUsed = list(ModelA,ModelB,ModelA,ModelB)

for (m in 1:4) {
  allpredictedCV = rep(NA,n)
  for (i in 1:nfolds)  {
    if (MethodsUsed[m] == "LDA") {
      modelfit = lda(ModelsUsed[[m]],subset=(cvgroups!=i), data=NewAuto)
    } else {
      modelfit = qda(ModelsUsed[[m]],subset=(cvgroups!=i), data=NewAuto)
    }
    newdata <- NewAuto[cvgroups==i,]
    allpredictedCV[cvgroups==i] = predict(modelfit,newdata)$class
  }
  CVmodels1_4[m] = sum(allpredictedCV!=NewAuto$origin)/n
}
CVmodels1_4
```

Now, we compare Models 2 and 4. Determine the number of parameters that must be estimated for each model.

Possible Code Answer:

```
K=3; p=1; K+K*p+p*(p+1)/2  # Model 1
K=3; p=5; K+K*p+p*(p+1)/2  # Model 2
K=3; p=1; K+K*p+K*p*(p+1)/2  # Model 3
K=3; p=5; K+K*p+K*p*(p+1)/2  # Model 4
```

Question

Identify which you would prefer between model 2 (LDA) and model 4 (QDA). Discuss the $CV_{(10)}$ values in light of: the number of parameters that need to be estimated and a comparison of the underlying assumptions about predictor variability for each model.

Possible Answer: *Despite the clear lack of constant variance, the large number of parameters needed for estimating the covariance matrices for QDA results in a similarly-accurate model for prediction. I would opt for the simpler model (2), barring other information.*

From Problem 4: Checking Assumptions

Question

For the models (LDA and QDA) using all five predictors, check the assumption of multivariate normality for the three sets of predictor variables (split by *origin*). Enter your R commands below.

Possible Code Answer:

```
names(NewAuto)
xvar = NewAuto %>%
  select(mpg, cylinders, displacement, horsepower, weight)
# Xmatrix within each class
xAmerican = NewAuto %>%
  filter(origin==1) %>%
  select(mpg, cylinders, displacement, horsepower, weight)
xEuropean = NewAuto %>%
  filter(origin==2) %>%
  select(mpg, cylinders, displacement, horsepower, weight)
xJapanese = NewAuto %>%
  filter(origin==3) %>%
  select(mpg, cylinders, displacement, horsepower, weight)

# check for multivariate normality
mhz(xAmerican)$mv.test
mhz(xEuropean)$mv.test
mhz(xJapanese)$mv.test
```

Question

Provide an alternative method, suited to the qualitative response *origin*, that could be used to fit the above model. You may provide explanation / reasoning to support your choice.

Possible Answer: *We could use k-nearest neighbors, as it has minimal assumptions about data distribution. (note that standardization of data would be necessary).*