Homework 6

Snigdha Peddi

Question 1: Question 4.7.7 pg 170 show your work, feel free to use R and use echo = T to show your code.

Density Function:

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

For the companies that did not issue dividend,$\mu = 0$:

By plugging in the given values $x = 4, \sigma^2 = 36$ in the density function we get,

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi * 36}} e^{\frac{-(4-0)^2}{2*36}} \\ &= \frac{1}{6\sqrt{2\pi}} e^{\frac{-16}{2*36}} \\ &= \frac{1}{6\sqrt{2\pi}} e^{\frac{-2}{9}} \end{aligned}$$

$$= \frac{1}{6 * 2.5066282746}(0.8007374029)$$
$$= 0.0532413343$$

For the companies that issued dividend, $\mu = 10$:

By plugging in the given values $x = 4, \sigma^2 = 36$ in the density function we get,

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi * 36}} e^{\frac{-(4-10)^2}{2*36}} \\ &= \frac{1}{6\sqrt{2\pi}} e^{\frac{-36}{2*36}} \\ &= \frac{1}{6\sqrt{2\pi}} e^{\frac{-1}{2}} \end{aligned}$$

$$= \frac{1}{6 * 2.5066282746}(0.6065306597)$$

$$= 0.0403284541$$

According to Bayes theorem,

$$
\begin{aligned}
P_{(D=Yes/f(4))} &= \frac{(0.0403284541)(0.80)}{(0.0532413343)(0.20)+(0.0403284541)(0.80)} \\
&= \frac{(0.0322627633)}{0.0106482669+0.0322627633} \\
&= \frac{(0.0322627633)}{0.0429110302} \\
&= 0.7518524526
\end{aligned}
$$

The probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year is 75.18% .

Question 2:Continue from Homework #3 & 4 using the Auto dataset from 4.7.11. Construct a model (using the predictors chosen for previous homework) and fit this model using MclustDA function from the mclust library. Use the same training and test set from previous homework assignments.

```
##
## Size of Training Data: 274
```

```
## Size of Test Data: 118
```

i) Provide a summary of your model. • What is the best model using BIC as the model selection criteria? Report the model name and BIC. (See mclustModelNames) • Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework # 3.

I have used the predictors cylinders,horsepower,weight,displacement to fit my discriminant model using *MclustDA()* function and Training data. 8 models were fit with different cluster groups ranging from G=1 to G=8.

BIC of all the models is compared to find the best model.The mixture model with G=3 has a largest BIC compared to other models(-8945.465).

Table 1: Comparision of BIC for different number of Clusters

G1.BIC	G2.BIC	G3.BIC	G4.BIC	G5.BIC	G6.BIC	G7.BIC	G8.BIC
-9864.135	-9226.066	-8945.465	-9046.725	-9104.819	-9183.838	-9282.565	-9184.923

Table 2: Comparision of Training Error for different number of Clusters

G1.Error	G2.Error	G3.Error	G4.Error	G5.Error	G6.Error	G7.Error	G8.Error
0.1058394	0.0766423	0.080292	0.080292	0.080292	0.080292	0.080292	0.0729927

The model with the best BIC is EEV model.It is of ellipsoidal,equal volume and equal shape.

Table 3: Model and Group

Class	Model	Group
0	EEV	3
1	EEV	3

```
##
## Mixture Model with best BIC is EEV:ellipsoidal,equal volume and equal shape
```

Table 4: Other Metrics of model

	n	Proportion	Brier Score	loglike
0	143	0.52	0.069	-4270.66
1	131	0.48	0.069	-4270.66

However, when a mixture model is fit using G=1:9 groups $(mod.DA.G19 < -MclustDA(x.dat, class.dat, G = 1 : 9))$ the BIC is -8871.28 which larger than fitting the models individually. Though the BIC with 3 clusters groups is close to this model(-8945.465) the training error rate of model with 1:9 groups is lower,6.5% compared to 8.0%.

```
## Training Error of model where G=1:9: 6.569343 %
```

```
## BIC of the Model where G=1:9: -8871.28
```

Table 5: Model and Group

Class	Model	Group
0	EEV	8
1	EEV	2

```
##
## Mixture Model with best BIC is EEV:ellipsoidal,equal volume and equal shape
```

Table 6: Other Metrics of model

	n	Proportion	Brier Score	loglike
0	143	0.52	0.06	-4110.079
1	131	0.48	0.06	-4110.079

In both cases the best model selected is EEV (EEV-ellipsoidal,equal volume and equal shape).

Below are the metrics for model with 3 clusters,

```
mod.DA.G3<-MclustDA(x.dat,class.dat,G=3)
```

Table 7: Metrics for Training Data

Metrics	Values
TPR	0.924
TNR	0.916
Training Error	0.080

Table 8: Metrics for Test Data

	Values
TPR	0.908
TNR	0.887
Test Error	0.102

Below are the metrics for mixture model with 1 to 9 clusters,

```
mod.DA.G19<-MclustDA(x.dat,class.dat,G=1:9)
```

Table 9: Metrics for Training Data

Metrics	Values
TPR	0.954
TNR	0.916
Training Error	0.066

Table 10: Metrices for Test Data

	Values
TPR	0.923
TNR	0.906
Test Error	0.085

The mixture model with G=1:9, has a lower training error of 6.6% and test error of 8.5% compared to model with G=3, where training error of 8.0% and test error of 10.2%.

ii) Specify modelType = "EDDA" and run MclustDA again. Provide a summary of your model. • What is the best model using BIC as the model selection criteria? Report the model name and BIC. • Report the true positive rate, true negative rate, training error, and test error.

I have used the same predictors cylinders,horsepower,weight,displacement to fit my discriminant model using *MclustDA()* function with modelType = "EDDA (same covarient matrix used for all clusters) and Training data.

```
mod.DA.Edda<-MclustDA(x.dat,class.dat,modelType = "EDDA")
```

```
## Training Error of EDDA model: 10.58394 %
```

```
## BIC of the EDDA Model : -9863.803
## EDDA Model with best BIC is VEV: ellipsoidal,equal shape
```

Table 11: Model and Group

Class	Model	Group
0	VEV	1
1	VEV	1

Table 12: Other Metrics of EDDA model

	n	Proportion	Brier Score	loglike
0	143	0.52	0.093	-4861.737
1	131	0.48	0.093	-4861.737

Table 13: Metrics for Training Data(EDDA)

Metrics	Values
TPR	0.901
TNR	0.888
Training Error	0.106

Table 14: Metrices for Test Data(EDDA)

	Values
TPR	0.923
TNR	0.849
Test Error	0.110

EDDA Model has a VEV, ellipsoidal,equal shape. The mixture model has a training error of 10.6% and test error of 11.0%.

iii)Compare the results with Homework #3 & 4. Which method performed the best? Justify your answer. Present your results in a well formatted table; include the previous methods and their corresponding rates.

A Logistic regression model is fit for training data using the variables cylinders,displacement,horsepower,weight as predictors and mpg01 variable as response variable.

```
reg.mod3<-glm(mpg01~cylinders+displacement+horsepower+weight,
          data=train1, family=binomial)
```

Using the same predictors and response variable as before a LDA model is fit

```
auto.lda <- lda(mpg01~cylinders+displacement+horsepower+weight,data=train1)
```

Using the same predictors and response variable as before a QDA model is fit.

```
auto.qda <- qda(mpg01~cylinders+displacement+horsepower+weight,data=train1)
```

Using the same predictors and response variable as before a knn model is fit.The graph shows the misclassification, sensitivity, and specificity for k values from 1 to 100 were investigated and k value of 2 is picked for the analysis beacuse of lower misclassification rate and a reasonable sensitivity and specificity.

Using the same predictors and response variable MclustDA and MclustDA_Edda models were fit (Question 2.i,2.ii)

Below is the table showing the comparative metrics from all the models on Auto data

Table 15: Comparision Metrics of Auto Data

	GLM	LDA	QDA	knn	MclustDA	Mclust_EDDA
TPR	0.923	0.969	0.908	0.371	0.923	0.923
TNR	0.849	0.811	0.868	0.562	0.906	0.849
Test Error	0.110	0.102	0.110	0.511	0.085	0.110

From the comparative results it is clear that MclustDA model yield better results with lower misclassification rate of 8.5%, high Sensitivity(92.3%)and high specificity(90.6%).

iv) From the original model variables, construct a new set of variables, fit a model using MclustDA and repeat i-iii. Hint: new variables may be interactions, polynomials, and/or splines. Do these new variables give an improvement in error rates compared to previous models? Explain how the new variables were constructed

New polynomial variables of all the predictors and few interaction terms were added to the original dataset. The data is divided into Train and test set by 70:30 split using the same seed (easy to compare results from 2 different models on same set of test and train records) used for initial analysis.This data is used to fit Model1. Due to few variables a high p value was observed for all terms.Model2 was fit by removing variable "name" and related polynomial terms. Model2 also resulted in high p values of all terms. Another model is fit removing all the polynomial terms

```
auto.glm.new2 <- glm(mpg01~.,data=train3[,c(1:8,21:24)],family='binomial')
```

This model resulted in few terms that are highly correlated like "acceleration","weight","year" and an interaction term between "weight" &"acceleration". A new model is fit with these variables and is used for the analysis.

```
auto.glm.new3 <- glm(mpg01~weight+acceleration+year+int_3,data=train3,family='binomial')
```

The same variables are used to fit logistic regression model, LDA model, QDA model,KNN Model(a k of 9 is used for analysis.A knn classification is performed with a k value of 1 to 100 and at k=9 the missclassification rate is at the lowest with reasonable sensitivity and specificity),MclustDA model and Mclust_EDDA model.

```
## Dimensions of New Auto dataset: 392 24

##
## Size of New Training Data: 274

## Size of New Test Data: 118
```

8 *MclustDA()* models were fit with different cluster groups ranging from G=1 to G=8 using the new training datset and previously selected variables. The tables below show the BIC and Training error rates of all these models.

The mixture model with 6 clusters has a higher BIC(-11363.64) compared to other models and I have used this model for further analysis and predicting the Test error,Sensitivity and Specificity.

```
mod.DA.G66<-MclustDA(x.dat1,class.dat1,G=6)
```

Table 16: Comparision of BIC for different number of Clusters

G1.BIC	G2.BIC	G3.BIC	G4.BIC	G5.BIC	G6.BIC	G7.BIC	G8.BIC
-11927.19	-11589.83	-11491.39	-11484.5	-11403.4	-11363.64	-11490.36	-11510.67

Table 17: Comparision of Training Error for different number of Clusters

G1.Error	G2.Error	G3.Error	G4.Error	G5.Error	G6.Error	G7.Error	G8.Error
0.0839416	0.0912409	0.0620438	0.0620438	0.0656934	0.0620438	0.040146	0.0364964

Table 18: Model and Group

Class	Model	Group
0	EEV	6
1	VEV	6

```
##
## Mixture Model with best BIC is EEV:ellipsoidal,equal volume and equal shape for class 0
##
##  and VEV:ellipsoidal equal shape for class 1
```

Table 19: Other Metrics of model

	n	Proportion	Brier Score	loglike
0	143	0.52	0.04	-5280.483
1	131	0.48	0.04	-5280.483

Mixture Model with 6 clusters with best BIC is *EEV*:ellipsoidal,equal volume and equal shape for class ‘0’(mpg lower than median value) and *VEV*:ellipsoidal equal shape for class ‘1’(mpg higher than median value).

Misclassification rate, specificity and sensitivity of the test data is calculated.

Table 20: Metrics for Training Data

Metrics	Values
TPR	0.954
TNR	0.923
Training Error	0.066

A new MclustDA model is fit with modelType as EDDA.

```
mod.DA.Edda1<-MclustDA(x.dat1,class.dat1,modelType = "EDDA")
```

EDDA Model with best BIC is VVV: ellipsoidal, varying volume, shape, and orientation.Other metrics were calculated and presented below.

```
## Training Error of EDDA model: 8.394161 %
```

```
## BIC of the EDDA Model : -11927.19
```

```
## EDDA Model with best BIC is VVV: ellipsoidal, varying volume, shape, and orientation
```

Table 21: Model and Group

Class	Model	Group
0	VVV	1
1	VVV	1

Table 22: Other Metrics of EDDA model

	n	Proportion	Brier Score	loglike
0	143	0.52	0.072	-5885.01
1	131	0.48	0.072	-5885.01

Table 23: Metrics for Training Data(EDDA)

Metrics	Values
TPR	0.947
TNR	0.888
Training Error	0.084

Table 24: Metrices for Test Data(EDDA)

	Values
TPR	0.923
TNR	0.849

	Values
Test Error	0.110

Below is the table showing the comparative metrics from all the models on Auto data with new set of variables.

Table 25: Comparision Metrics of Auto Data

	GLM	LDA	QDA	knn	MclustDA	Mclust_EDDA
TPR	0.892	0.938	0.923	0.448	0.831	0.923
TNR	0.906	0.849	0.849	0.536	0.849	0.849
Test Error	0.102	0.102	0.110	0.481	0.161	0.110

All the models were compared with the previously generated models and below are my observations,

- For Logistic regression model the new variables resulted in little lower test error, higher sensitivity and little lower specificity.
- For LDA model the new variables resulted in same test error, little lower sensitivity and little higher specificity.
- For QDA model the new variables resulted in same test error, higher sensitivity and little lower specificity.
- For knn model the new variables resulted in lower test error, higher sensitivity and little lower specificity.
- For MclustDA model the new variables resulted in significantly higher test error, lower sensitivity and lower specificity.
- For MclustDA model with EDDA modeltype the new variables resulted in same test error,sensitivity and specificity.

  In conclusion,comparing all the different models MclustDA model using the initial variables cylinders,weight,horsepower and displacement resulted in the better model with a lower test error of 8.5%, higher sensitivity of 92.3% and highr specificity of 90.6%.

REFERENCES

- Stat 602 Lecture,Introduction to Classification,*Playing With MclustDA part2* by Dr.Saunders.