Modern Applied Statistics exercises from ISLR

Yamuna Dhungana

1) Question 4.7.6 pg 170 Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta 0 = -6$, $\beta 1 = 0.05$, $\beta 2 = 1$.

a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

For the solution of this question, we simply plug the values of ˆ 0 , ˆ 1 and ˆ 2 in the equation:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1}}{(1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1})}$$

We have $\beta 0 = -6$,$\beta 1 = 0.05$ and $\beta 2 = 1$

$$\hat{p}(X) = \frac{e^{-6+0.05X_1+X_2}}{(1 + e^{-6+0.05X_1+X_2})} = \frac{e^{-6+0.05*40+1*3.5}}{(1 + e^{-6+0.05*40+1*3.5})} = 0.3775.$$

```
## We can also write a function in R to calculate probability
```

```
## [1] 0.3775
```

b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

The equation for predicted probability gives us:

$$\frac{e^{-6+0.05X_1+3.5}}{(1 + e^{-6+0.05X_1+3.5})} = 0.5$$

from which, we get:

$$e^{-6+0.05X_1+3.5} = 1$$

If we take log of both sides, we get:

$$X_1 = \frac{2.5}{0.05} = 50$$

```
## We can also use the function to calculate this and get the answer
```

```
## 40_hours 41_hours 42_hours 43_hours 44_hours 45_hours 46_hours 47_hours
##   0.3775   0.3894   0.4013   0.4134   0.4256   0.4378   0.4502   0.4626
## 48_hours 49_hours 50_hours 51_hours 52_hours 53_hours 54_hours 55_hours
##   0.4750   0.4875   0.5000   0.5125   0.5250   0.5374   0.5498   0.5622
## 56_hours 57_hours 58_hours 59_hours 60_hours
##   0.5744   0.5866   0.5987   0.6106   0.6225
```

Hence, to have 50 percent chance for securing A, the student has to study for 50 hours.

2) Continue from Homework #3 & 4 using the **Weekly** dataset from 4.7.10). Construct a model (using the predictors chosen for previous homework) and fit this model using the `MclustDA` function from the **mclust** library.

i) Provide a summary of your model.

- What is the best model selected by BIC? Report the Model Name and the BIC. (See[mclustModelNames] (https://www.rdocumentation.org/packages/mclust/versions/5.4/topics/mclustModelNames))
- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework # 3.

```
## ----------------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -2129.439 985 10 -4327.804
##
## Classes   n     % Model G
##    Down 441 44.77     V 2
##    Up   544 55.23     V 2
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   76 365
##   Up     70 474
## Classification error = 0.4416
## Brier score          = 0.2452
```

```
##          train_error
## accuracy       55.84
## TPR            87.13
## TNR            17.23
```

```
##          test_error
## accuracy      54.81
## TPR           85.25
## TNR           11.63
```

```
## ## With all the variables
```

```
## ----------------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n  df       BIC
##       -12477.54 985 286 -26926.37
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 4
##    Up   544 55.23   VVV 4
##
```

```
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  251 190
##   Up    160 384
## Classification error = 0.3553
## Brier score          = 0.2177
```

```
##          train_error
## accuracy       64.47
## TPR            70.59
## TNR            56.92
```

```
##          test_error
## accuracy      53.85
## TPR           85.25
## TNR            9.30
```

From the previous work, I have selected Lag2 as the most significant variable. However, I am performing two models, one with only Lag2 and the other with all the variables. I only want to see how all the variables work. From the single variable, the model is variable variance (one-dimensional) with the two groups. And for the model with all the variables, the model is ellipsoidal, varying volume, shape, and orientation. The BIC of the single variable model is more than the BIC with all the variables. I have also made the table for both the test and train for these models.

ii) Repeat the `MclustDA` analysis, but this time specify `modelType = "EDDA"`. Provide a summary of this model. * What is the best model using BIC as the model selection criteria?

```
* Report the true positive rate, true negative rate, training error, and test error. You can reuse the
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -15029.33 985 49 -30396.39
##
## Classes   n     % Model G
##    Down 441 44.77   VVE 1
##    Up   544 55.23   VVE 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   98 343
##   Up     90 454
## Classification error = 0.4396
## Brier score          = 0.2497
```

```
##          train_error
## accuracy       56.04
## TPR            83.46
## TNR            22.22
```

```
##          test_error
## accuracy      46.15
```

```
## TPR          14.75
## TNR          90.70
```

I have fitted the model with all the variables and with a single variable. However, the model with the single variable could not converge. I have made the table for the train and the test errors for the model. From the documentation of Mclust, it was found that specifying "EDDA" as the model type, we force the model to have a single component in each class with the same covariance structure. We see that the single component had an ellipsoidal, equal orientation (VVE) structure.

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.*

Table 1: MsclustDA Test Accuracy with single variable

	train_error	test_error
accuracy	55.84	54.81
TPR	87.13	85.25
TNR	17.23	11.63

Table 2: MsclustDA Test Accuracy with all variables

	train_error	test_error	train_error	test_error
accuracy	64.47	53.85	56.04	46.15
TPR	70.59	85.25	83.46	14.75
TNR	56.92	9.30	22.22	90.70

Table 3: MsclustDA with EDDA Test Accuracy with all variables

	train_error	test_error
accuracy	56.04	46.15
TPR	83.46	14.75
TNR	22.22	90.70

Table 4: Logreg Accuracy measures with single variable

	Train_error	Test_error
accuracy	55.53	62.50
TPR	96.32	91.80
TNR	5.22	20.93

Table 5: LDA Accuracy measures with single variable

	Train_error	Test_error
accuracy	55.43	62.50
TPR	96.32	91.80
TNR	4.99	20.93

Table 6: QDA Accuracy measures with single variable

	Train_error	Test_error
accuracy	55.23	58.65
TPR	100.00	100.00
TNR	0.00	0.00

Table 7: KNN Accuracy measures with single variable

	Test_error
accuracy	50.96
TPR	52.46
TNR	49.00

Here, I have made some tables for all the methods that we fitted on this as well as previous. By looking at the table, we can find out that the accuracy of the test data varies from 62.50 to 46.15. Likewise, training accuracy varies from 64.47 and 50.0. The accuracy of the logistic regression is highly accurate besides logistic regression, LDA also has a high accuracy rate.

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.

```
## ----------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -4220.728 985 18 -8565.523
##
## Classes   n     % Model G
##    Down 441 44.77   VII 3
##    Up   544 55.23   VII 2
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  156 285
##   Up    137 407
## Classification error = 0.4284
## Brier score          = 0.243
```

```
## ----------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
```

```
##       -5545.222 985 90 -11710.78
##
## Classes   n     % Model G
##    Down 441 44.77   VEV 5
##    Up   544 55.23   VVV 5
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  204 237
##   Up    200 344
## Classification error = 0.4437
## Brier score          = 0.2945
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -2314.625 985 58 -5029.024
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 5
##    Up   544 55.23   VVV 5
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  125 316
##   Up    128 416
## Classification error = 0.4508
## Brier score          = 0.3041
```

Table 8: Accuracy measures using MclustDA

	tr.error modd1	tt.error modd1	tr.error modd2	tt.error modd2	tr.error modd3	tt.error modd3
accuracy	57.16	50.96	55.63	55.77	54.92	57.69
TPR	74.82	70.49	63.24	60.66	76.47	77.05
TNR	35.37	23.26	46.26	48.84	28.34	30.23

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -4408.247 985  5 -8850.957
##
## Classes   n     % Model G
##    Down 441 44.77   EII 1
```

```
##    Up   544 55.23   EII 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   50 391
##   Up     52 492
## Classification error = 0.4497
## Brier score          = 0.2454
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -7896.142 985 18 -15916.35
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 1
##    Up   544 55.23   VVV 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down  311 130
##   Up    355 189
## Classification error = 0.4924
## Brier score          = 0.256
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -6144.173 985 10 -12357.27
##
## Classes   n     % Model G
##    Down 441 44.77   VVV 1
##    Up   544 55.23   VVV 1
##
## Training confusion matrix:
##       Predicted
## Class  Down  Up
##   Down   14 427
##   Up     19 525
## Classification error = 0.4528
## Brier score          = 0.2558
```

Table 9: Accuracy measures using MclustDA EDDA

	tr.error.1ed	tt.error.1ed	tr.error.2ed	tt.error.2ed	tr.error.3ed	tt.error.3ed
accuracy	55.03	56.73	50.76	46.15	54.72	62.50
TPR	90.44	83.61	34.74	40.98	96.51	95.08
TNR	11.34	18.60	70.52	53.49	3.17	16.28

Table 10: Accuracy measures using LDA

	Trn_err.lda1	Tt_err.lda1	Trn_err.lda2	Tt_err.lda2	Trn_err.lda3	Tt_err.lda3
accuracy	54.82	57.69	54.82	57.69	54.82	61.54
TPR	91.54	86.89	91.54	86.89	96.69	93.44
TNR	9.52	16.28	9.52	16.28	3.17	16.28

Table 11: Accuracy measures using QDA

	Trn_err.qda1	Tt_err.qda1	Trn_err.qda2	Tt_err.qda2	Trn_err.qda3	Tt_err.qda3
accuracy	55.33	55.77	50.76	46.15	54.72	62.50
TPR	84.93	83.61	34.74	40.98	96.51	95.08
TNR	18.82	16.28	70.52	53.49	3.17	16.28

I have created three models, `Direction~Lag1+Lag2`,`Direction~Lag1+Lag2+Lag1*Lag2` `Direction~Lag2+I(Lag2^2)` for the I have fitted these three models with all the previous models that we have learned. Both models with MclustDA have a model spherical, unequal volume with the 3 groups for down and two for the up. Likewise, the second model has ellipsoidal, equal shape for down and Up with the ellipsoidal, varying volume, shape, and orientation with 5 groups. Additionally, the third model is ellipsoidal, varying volume, shape, and orientation with 5 groups. For the first model of EDDA, there is one group with a spherical, equal volume model. The second and third models have ellipsoidal, varying volume, shape, and orientation models with one group.I have made the table for all the model's accuracy in a combined table that represents the test and the training error.