

# Modern Applied Statistics exercises from ISLR

Yamuna Dhungana

- 1) Question 4.7.7 pg 170 *show your work, feel free to use R and use `echo = T` to show your code.* Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based on  $X$ , last year’s percent profit. We examine a large number of companies and discover that the mean value of  $X$  for companies that issued a dividend was  $\bar{X} = 10$ , while the mean for those that didn’t was  $\bar{X} = 0$ . In addition, the variance of  $X$  for these two sets of companies was  $\sigma^2 = 36$ . Finally, 80% of companies issued dividends. Assuming that  $X$  follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was  $X = 4$  last year. Hint: Recall that the density function for a normal random variable is

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

. You will need to use Bayes’ theorem.

**Solution** The values given in question are

$$\pi_{YES} = 0.8$$

$$\pi_{NO} = 0.2$$

$$\mu_{YES} = 10$$

$$\mu_{NO} = 0$$

$$\sigma^2 = 36$$

plugging the given values to the density function here,  $\pi_k$  is 0.8 and 0.2 and dividend is Yes and No.

So, using the density function to calculate  $f_k(x)$

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

we get,

$$f_{yes}(x) = 0.0402$$

$$f_{No}(x) = 0.0532$$

Form the ISLR book, using the equation 4.10 of Bayes Theorem

$$P_{(divident=K|X=x)} = \frac{\pi_k f_k(x)}{\sum_{l=1}^k \pi_l f_l(x)}$$

again, plunging the value we calculated earlier in the above equation.

$$P_{yes}(4) = \frac{0.8 \times 0.04032}{0.8 \times 0.0402 + 0.2 \times 0.0532} = 0.75186$$

**The same can be done in R and get the result:**

```

# pdf function
pdf <- function(x, mu_k, sigma){((sqrt(2*pi)*sigma)^-1)*(exp(-((2*sigma^2)^-1)*(x-mu_k)^2))}

sigma <- 6 # for both classes

# Type 1 where companies issued dividend
pi_1 <- .8
mu_1 <- 10

# Type 2 where companies did not issue dividend
pi_2 <- .2
mu_2 <- 0

# Calculate probabilities based on Bayes
x <- 4
p_1 <- (pi_1*pdf(4,mu_1,sigma))/(pi_1*pdf(4,mu_1,sigma) + pi_2*pdf(4,mu_2,sigma))
p_2 <- (pi_2*pdf(4,mu_2,sigma))/(pi_1*pdf(4,mu_1,sigma) + pi_2*pdf(4,mu_2,sigma))

# rounding the numbers
p_1 <- round(p_1,4)
p_2 <- round(p_2,4)

# prediction
prediction <- data.frame(cbind(c("Dividend", "Non-Dividend"), c(p_1, p_2)))
colnames(prediction) <- c("Types", "prediction")
prediction

```

```

##           Types prediction
## 1      Dividend      0.7519
## 2 Non-Dividend      0.2481

```

Based on my analysis 75.19 % of the companies will announce the dividend this year and 24.81% will not do so.

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.

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] (<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
##      -4393.443 245 215 -9969.657
##
## Classes      n      % Model G

```

```
##      0 125 51.02  EEV 3
##      1 120 48.98  EEV 4
##
## Training confusion matrix:
##      Predicted
## Class    0    1
##      0 115   10
##      1   2  118
## Classification error = 0.049
## Brier score          = 0.0497
```

Table 1: Best model selected by BIC

Model names	BIC
EEV	-9969.65745670221
EEV	-9969.65745670221

Like the previous work, I have split the data into test and train sets. I have divided the dataset into a ratio of 70 to 30. I have fitted the MclustDA model. Following the previous work, the response was the binary coded MPG (whether greater or equal to median mpg). Same predictors as previous were used as previous based on their correlation with mpg01. The BIC of the model is -10568.327944618 and has the model ellipsoidal, equal volume, and equal shape with the 3 and 4 groups.

Table 2: Mclust model accuracy measures

	Mclust train set	Mclust test set
accuracy	95.10	88.44
TPR	98.33	89.47
TNR	92.00	87.32

Using the same function as in homework 3 and 4. I used the same function to find out the test accuracies of the model. The accuracy of the train and the test model is presented in table-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.

```
## -----
## Gaussian finite mixture model for classification
## -----
##
## EDDA model summary:
##
## log-likelihood  n df      BIC
##      -5525.729 245 70 -11436.55
##
## Classes    n      % Model G
##      0 125 51.02  VVV 1
##      1 120 48.98  VVV 1
##
## Training confusion matrix:
##      Predicted
## Class    0    1
```

```
##      0 112 13
##      1   6 114
## Classification error = 0.0776
## Brier score          = 0.0652
```

Table 3: Best model selected by BIC with EDDA

Model names	BIC
VVV	-11436.5461952401
VVV	-11436.5461952401

Fitting the MclustDA model with the model type EDDA we get the BIC with -11500.33 with the model name ellipsoidal, varying volume, shape, and orientation (VVV), and has a single group.

Table 4: Mclust with EDDA model accuracy measures

	Mclust:EDDA train set	Mclust:EDDA test set
accuracy	92.24	90.48
TPR	95.00	89.47
TNR	89.60	91.55

The Model accuracies are reported in the table-4

- 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 5: Mclust models accuracy measures

	Mclust train set	Mclust test set	Mclust:EDDA train set	Mclust:EDDA test set
accuracy	95.10	88.44	92.24	90.48
TPR	98.33	89.47	95.00	89.47
TNR	92.00	87.32	89.60	91.55

Table 6: Logreg Accuracy measures

	Logreg.model Train set	Logreg.model Test set
accuracy	90.20	88.44
TPR	93.33	89.47
TNR	87.20	87.32

Table 7: LDA model Accuracy measures

	LDA.model Train set	LDA.model Test set
accuracy	91.43	89.80
TPR	95.83	93.42
TNR	87.20	85.92

Table 8: QDA model Accuracy measures

	QDA.model Train set	QDA.model Test set
accuracy	90.61	88.44
TPR	91.67	88.16
TNR	89.60	88.73

We have performed Logistic regression, LDA, QDA in the previous assignment, and in the present assignment I have performed MclustDA and with the model type EDDA. By looking at the table presented in table-5,6 7 and 8. We see that the accuracy of the MclustDA looks better than the other models and LDA and QDA have the least accuracy. Likewise, TPR and TNR also looks better in the MclustDA model.

- 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
##      -5014.757 245 90 -10524.63
##
## Classes      n      % Model G
##      0 125 51.02   VVV 5
##      1 120 48.98   VEV 5
##
## Training confusion matrix:
##      Predicted
## Class  0   1
##      0 109 16
##      1  13 107
## Classification error = 0.1184
## Brier score          = 0.0861
## -----
## Gaussian finite mixture model for classification
## -----
##
## MclustDA model summary:
##
##   log-likelihood    n df         BIC
##      -6371.577 245 35 -12935.7
##
## Classes      n      % Model G
##      0 125 51.02   VVE 3
##      1 120 48.98   EVV 4
##
## Training confusion matrix:
##      Predicted
## Class  0   1
```

```
##      0 107 18
##      1   7 113
## Classification error = 0.102
## Brier score          = 0.0738
```

Table 9: Best model selected by BIC with two models

Model1:model names	Model1:BIC	Model2:model names	Model2:BIC
VVV	-10524.6280312072	VVE	-12935.6971844703
VEV	-10524.6280312072	EVV	-12935.6971844703

Table 10: MclustDA:Accuracy measures for two models

	Mod1.mclustDA train set	mod1.mclustDA test set	Mod2.mclustDA train set	Mod2.mclustDA test set
accuracy	88.16	89.12	89.80	87.07
TPR	89.17	86.84	94.17	85.53
TNR	87.20	91.55	85.60	88.73

```
## -----
## Gaussian finite mixture model for classification
## -----
##
## EDDA model summary:
##
##   log-likelihood   n df      BIC
##      -5619.77 245 18 -11338.56
##
## Classes    n    % Model G
##      0 125 51.02   VVV 1
##      1 120 48.98   VVV 1
##
## Training confusion matrix:
##      Predicted
## Class  0   1
##      0  98 27
##      1   6 114
## Classification error = 0.1347
## Brier score          = 0.105
## -----
## Gaussian finite mixture model for classification
## -----
##
## EDDA model summary:
##
##   log-likelihood   n df      BIC
##      -6447.904 245 10 -12950.82
##
## Classes    n    % Model G
##      0 125 51.02   VVV 1
##      1 120 48.98   VVV 1
```

```
##
## Training confusion matrix:
##      Predicted
## Class    0    1
##      0  99  26
##      1   7 113
## Classification error = 0.1347
## Brier score          = 0.1021
```

Table 11: Best model selected by BIC with two models using EDDA

Model1:model names	Model1:BIC	Model2:model names	Model2:BIC
VVV	-11338.5626652424	VVV	-12950.8195845287
VVV	-11338.5626652424	VVV	-12950.8195845287

Table 12: MclustDA with EDDA:Accuracy measures for two models

	Mod1.edda train set	mod1.edda test set	Mod2.edda train set	mod2.edda test set
accuracy	86.53	85.03	86.53	82.99
TPR	95.00	90.79	94.17	86.84
TNR	78.40	78.87	79.20	78.87

```
##
## Call:
## glm(formula = formula1, family = binomial, data = newdata.train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33232  -0.24001  -0.02009   0.32681   3.02917
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.779e+01  7.712e+00   2.307   0.0211 *
## weight         -4.909e-03  2.574e-03  -1.907   0.0565 .
## horsepower     -8.413e-02  8.291e-02  -1.015   0.3102
## weight:horsepower 1.496e-05  2.702e-05   0.554   0.5798
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 339.54  on 244  degrees of freedom
## Residual deviance: 132.82  on 241  degrees of freedom
## AIC: 140.82
##
## Number of Fisher Scoring iterations: 8
##
## Call:
## glm(formula = formula2, family = binomial, data = newdata.train)
##
```

```
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3444  -0.2279  -0.0040   0.3367   3.4186
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.249e+01  9.198e+00   1.358  0.1744
## poly(weight, 2, raw = TRUE)1  1.212e-03  8.038e-03   0.151  0.8802
## poly(weight, 2, raw = TRUE)2 -8.611e-07  1.442e-06  -0.597  0.5505
## poly(horsepower, 2, raw = TRUE)1 -1.458e-01  7.987e-02  -1.825  0.0679 .
## poly(horsepower, 2, raw = TRUE)2  5.289e-04  3.686e-04   1.435  0.1513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 339.54  on 244  degrees of freedom
## Residual deviance: 131.85  on 240  degrees of freedom
## AIC: 141.85
##
## Number of Fisher Scoring iterations: 8
```

Table 13: Logistic regression: Accuracy measures for two models

	mod1.Logreg Train set	mod1.Logreg Test set	mod2.Logreg Train set	mod2.Logreg Test set
accuracy	87.35	89.80	87.76	89.12
TPR	87.50	85.53	88.33	85.53
TNR	87.20	94.37	87.20	92.96

Table 14: LDA: Accuracy measures for two models

	LDA.mod1 Train set	LDA.mod1 Test set	LDA.mod2 Train set	LDA.mod1 Test set
accuracy	87.76	89.80	88.16	89.80
TPR	87.50	85.53	87.50	85.53
TNR	88.00	94.37	88.80	94.37

Table 15: QDA: Accuracy measures for two models

	LDA.mod1 Train set	QDA.mod1 Test set	QDA.mod2 Trainset	QDA.mod1 Test set
accuracy	86.53	85.03	84.90	83.67
TPR	95.00	90.79	96.67	90.79
TNR	78.40	78.87	73.60	76.06

Previously, I have found that cylinders, weight, displacement, and horsepower are mostly associated with the mpg01 variable. However, in the logistic regression, I found out that the P-values for weight and horsepower are statistically significant. Hence, I only used these two variables for the interaction and as the polynomial. The interaction term is `mpg01 ~ weight + horsepower + horsepower:weight` and the polynomial is `mpg01 ~ poly(weight, 2, raw = TRUE) + poly(horsepower, 2, raw = TRUE)`. I have



fitted MclustDA, MclustDA with EDDA, Logistic regression, LDA, and QDA models with these variables. From the MclustDA and Mclust with EDDA, the BIC of models are presented in table- 9 and table-11. Model-1 has a higher BIC than the second model and is considered as the better fit. The Model name of model-1 are ellipsoidal, equal volume and equal shape (EEV), and ellipsoidal, equal shape (VEV). Likewise, The model name for the second model are ellipsoidal, equal orientation (new models in mclust version  $\geq 5.0.0$ )(VVE) and ellipsoidal, equal orientation (new models in mclust version  $\geq 5.0.0$ )(VEE). For the MclustDA with EDDA, model-1 also has the highest BIC than the second model. Both the models have the model name same which is ellipsoidal, varying volume, shape, and orientation (VVV). I have Fitted all the models previously solved with the same interaction, and polynomial terms. I have calculated the accuracy measures. The accuracy of the models are presented in the table above. The accuracy of the model varies from 89.9 to 82.99. MclustDA model has the highest accuracy. The accuracy of the model does not improve with these new variables.