Homework 6

Rylie Fleckenstein

2/24/21

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 $\hat{\sigma}^2 = 36$. 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

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

You will need to use Bayes' theorem.

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}}{\sum_{i=1}^{k} \pi_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_i)^2}}$$

We know that $\mu_1 = 10$, $\mu_2 = 0$, $\sigma^2 = 360$, $\pi_1 = 0.8$, $\pi_2 = 0.2$

$$p(4) = \frac{0.8 * e^{-\frac{1}{2*36}*(4-10)^2}}{0.8 * e^{-\frac{1}{2*36}*(4-10)^2} + 0.2 * e^{-\frac{1}{2*36}*(4-0)^2}}$$

$$p(4) = \frac{0.8 * e^{-\frac{1}{2}}}{0.8 * e^{-\frac{1}{2}} + 0.2 * e^{-\frac{2}{9}}}$$

$$p(4) = 0.752$$

Therefore, given a company with a percentage profit of X=4 for last year, the probability of said company issuing a dividend for this year is at 75.2%.

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.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -4037.774 274 72 -8479.694
##
## Classes   n  % Model G
##       0 137 50   EEV 3
##       1 137 50   EEV 3
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 127  10
##     1  11 126
## Classification error = 0.0766
## Brier score          = 0.0638
```

- What is the best model using BIC as the model selection criteria? Report the model name and BIC. (See mclustModelNames)

The best model using BIC is the EEV model. According to the documentation the EEV model stands for "ellipsoidal, equal volume and equal shape" and the BIC for this model is -8479.69.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework # 3.

From taking a look at the model summary we can see th training error was around 7.66%, and then below we can see the test error comes in at 11.86%, the true positive rate is 86.44%, and the true negative rate is 89.83% giving us an overall accuracy of 88.14%.

```
## $`Confusion Matrix`
##
##      0  1
##   0 51  8
##   1  6 53
##
## $`Overall Percentage Correct`
## [1] 88.13559
##
## $Sensitivity
## [1] 86.44068
##
## $Specificity
## [1] 89.83051
##
## $`Test Error`
## [1] 11.86441
```

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.

After changing modelType="EDDA" and running the same analysis we get the results as shown below. We can see that the best model using BIC is the "VEV" mode which stands for "ellipsoidal, equal shape" and the BIC for this model is -9709.27. If we were to compare this model with the previous one based upon BIC alone, we could say that this model is a better fit to our data because the BIC value is lower.

The training error for this model is 8.76%, the test error is 14.40%, the true positive rate is 84.75%, the true negative rate is 86.44%, giving us an overall accuracy of 85.59%. Based upon these metrics we can say that this model performed worse than the first model in this analysis.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -4784.472 274 25 -9709.272
##
## Classes   n  % Model G
##       0 137 50   VEV 1
##       1 137 50   VEV 1
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 124  13
##     1  11 126
## Classification error = 0.0876
## Brier score          = 0.0778
```

```
## $`Confusion Matrix`
##
##      0  1
##   0 50  9
##   1  8 51
##
## $`Overall Percentage Correct`
## [1] 85.59322
##
## $Sensitivity
## [1] 84.74576
##
## $Specificity
## [1] 86.44068
##
## $`Test Error`
## [1] 14.40678
```

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.*

If we take a look at the results from homework 3,4, and now 6 we can can see all models performed pretty well. The best two models for this problem were the LDA and MclustDA models. Both of them have a test error of 11.86 and an accuracy of 88.14%. With that being said, if we take a look at the true positive and true negative rates for the two models we can see the MclustDA model has a more evenly distributed prediction accuracy among the two classes where the LDA is slightly more skewed. Because of this, I would have to chose the MclustDA model as the best model in this comparison.

Table 1: Model Scores

Models	Accuracy	Sensitivity	Specificity	Test.Error
Log Reg	86.44	83.05	89.83	13.56
LDA	88.14	81.36	94.92	11.86
QDA	85.59	84.75	86.44	14.41
KNN	84.75	90.20	80.60	15.25
MclustDA	88.14	86.44	89.83	11.86

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.

The new model created with variable interactions and transformations does in fact give an improvement in error rates compared to the previous models. I was able to find a combination of variable transformations that resulted in an improvement and those transformations are the following: an interaction between cylinders*weight, squaring horsepower^2, and squaring displacement^2. I created a variable matrix containing these transformations for the training set and the test set below which allowed me to use them in a MclustDA model.

i) Provide a summary of your model.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log-likelihood   n df       BIC
##       -5355.994 274 42 -10947.74
##
## Classes   n  % Model G
##       0 137 50   VVE 3
##       1 137 50   VEE 3
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 129   8
##     1   9 128
## Classification error = 0.062
## Brier score          = 0.0511
```

- What is the best model selected by BIC? Report the Model Name and the BIC. (See mclustModelNames)

The best model selected by BIC for group 0 was the “VVE” which stands for “ellipsoidal, equal orientation” and the model selected for group 1 was “VEE” standing for “ellipsoidal, equal shape and orientation”. The BIC for this model is -10947.74.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework 3.

The training error for this model was 6.2% with a test error of 10.17%, true positive rate of 88.14%, and true negative rate of 91.53%, giving us an accuracy of 89.83%.

```
## $‘Confusion Matrix‘
##
##      0  1
##   0 52  7
##   1  5 54
##
## $‘Overall Percentage Correct‘
## [1] 89.83051
##
## $Sensitivity
## [1] 88.13559
##
## $Specificity
## [1] 91.52542
##
## $‘Test Error‘
## [1] 10.16949
```

ii) Repeat the `MclustDA` analysis, but this time specify `modelType = "EDDA"`. Provide a summary of this model.

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log-likelihood   n df       BIC
##       -5545.668 274 18 -11192.37
##
## Classes   n  % Model G
##       0 137 50   VVV 1
##       1 137 50   VVV 1
##
## Training confusion matrix:
##      Predicted
## Class   0   1
##     0 122  15
##     1   9 128
## Classification error = 0.0876
## Brier score          = 0.0765
```

- What is the best model using BIC as the model selection criteria?

The model chose using BIC was the “VVV” model which stands for “ellipsoidal, varying volume, shape, and orientation”. The BIC for this model is -11192.37. So, if we were to chose a model based upon the BIC we would chose this model over the previous because this model’s BIC is lower making it a better fit model.

- Report the true positive rate, true negative rate, training error, and test error. You can reuse the function written in Homework 3.

The training error for this model is 8.76%, the test error is at 13.56%, true positive 83.05%, true negative 89.83%, and the accuracy for this model is 86.44%. Therefore, this model does not perform better than the first model and we will go with the first model as our chosen MclustDA model in this analysis.

```
## $‘Confusion Matrix‘
##
##      0  1
##   0 49 10
##   1  6 53
##
## $‘Overall Percentage Correct‘
## [1] 86.44068
##
## $Sensitivity
## [1] 83.05085
##
## $Specificity
## [1] 89.83051
##
## $‘Test Error‘
## [1] 13.55932
```

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.*

Taking a look at the test error rates we were able to achieve the lowest test error with the MclustDA model (10.17) that was fit with the various variable interactions outlined above. The next best model being the LDA model with a test error rate of 11.86.

Table 2: Model Scores

Models	Accuracy	Sensitivity	Specificity	Test.Error
Log Reg	86.44	83.05	89.83	13.56
LDA	88.14	81.36	94.92	11.86
QDA	85.59	84.75	86.44	14.41
KNN	84.75	90.20	80.60	15.25
MclustDA	89.83	88.14	91.53	10.17

Sources:

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer