STAT 602 - Homework 5

Snigdha Peddi in collaboration with John Herbert

Document External Libraries

- **ISLR** package for the homework data sets
- **knitr** package used for *kable* function used to format tables
- **dplyr** package for data formatting and cleaning
- **MASS** package for LDA and QDA models
- **mclust** package for Mclust models and graphs
- **class** package for KNN models

Reusable Functions

- The misclassification function created in homework 3 (*misclass.fun.JH*) will be reused for questions in this homework.

Exercises

Question 1 (ISLR 4.7.6 pg 170)

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

Question 1(a)

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

Answer

```
# Setting variables for the beta values for each each covariant 'x' and the intercept
b_0 = -6
b_1 = 0.05
b_2 = 1
# Setting variables for X values
x_1 = 40
x_2 = 3.5
# Calculating the "x" in phat_x
phat_x <- b_0+b_1*x_1+b_2*x_2
# Plug phat_x into our probability function to derive an answer
```

```r
phat <- exp(phat_x)/(1+exp(phat_x))
cat('Probability that this student will get an A is',round(phat*100,2),'%')
```

```
## Probability that this student will get an A is 37.75 %
```

In formulaic terms:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}$$

Plugging in values to the variables:

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

Simplified:

$$\hat{p}(X) = \frac{e^{-0.5}}{1+e^{-0.5}}$$

Therefore:

$$\hat{p}(X) = 0.3775407$$

Question 1(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?

Answer

Solve for x:

$$0.5 = \frac{e^{-6+0.05x+3.5}}{1+e^{-6+0.05x+3.5}}$$

Simplify the exponent:

$$0.5 = \frac{e^{-2.5+0.05x}}{1+e^{-2.5+0.05x}}$$

Multiply each side by the demoninator:

$$0.5(1+e^{-2.5+0.05x}) = e^{-2.5+0.05x}$$

Simplify the left side of the equation:

$$0.5+0.5(e^{-2.5+0.05x}) = e^{-2.5+0.05x}$$

Subtract each side by $0.5(e^{-2.5+0.05x})$:

$$0.5 = e^{-2.5+0.05x} - 0.5e^{-2.5+0.05x}$$

$$0.5 = 0.5e^{-2.5+0.05x}$$

Multiplying each side by 2:

$$1 = e^{-2.5+0.05x}$$

$$0 = -2.5 + 0.05x$$

Take the natural log of each side:

$$ln(1) = ln(e^{-2.5+0.05x})$$

$$0 = -2.5 + 0.05x$$

$$x = 50$$

It would take this student 50 hours of studying to have a 50% chance of recieving a 4.0 GPA.

Source: An Introduction to Statistical Learning

Question 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.

Question 2(i) Part I

Provide a summary of your model.

Answer

Using the all the *Lag* variables as was done in Homework 3 & 4 as the predictors, and *Direction* as the target with the *MclustDA* function to create the model. In addition, I set the # of Groups to 9. In order to calculate the training and test errors, I partitioned the data set into train/test with the same method as prior homeworks. The below summaires are just on the training set.

Table 1: Summary of Mclust Model (Train)

	Down	Up
n	441	544
Proportion	0.45	0.55
Model	VII	VII
G	3	2

Table 2: Metric Summary of Mclust Model (Train)

Metrics	Values
Class Error	0.424
Brier Score	0.242

Metrics	Values
Log Like	-10473.330
BIC	-21174.117

Table 3: Weekly MclustDA Confusion Matrix

	Pred Down	Pred Up
Act Down	185	256
Act Up	162	382

Question 2(i) Part II

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

Answer

Using the *Mclust* function, I examined the different BIC scores of groupings set betwen 1-9 for both the Up and Down class.

Table 4: Clustering Table Proportions for Up Class

Group	Proportion
1	0.817
2	0.183
3	0.817

Table 5: Model Metrics for Up Class

Metrics	Values
Model	VII
Log Like	-5795.5
n	544
df	5
BIC	-11672.89
ICL	-11743.48

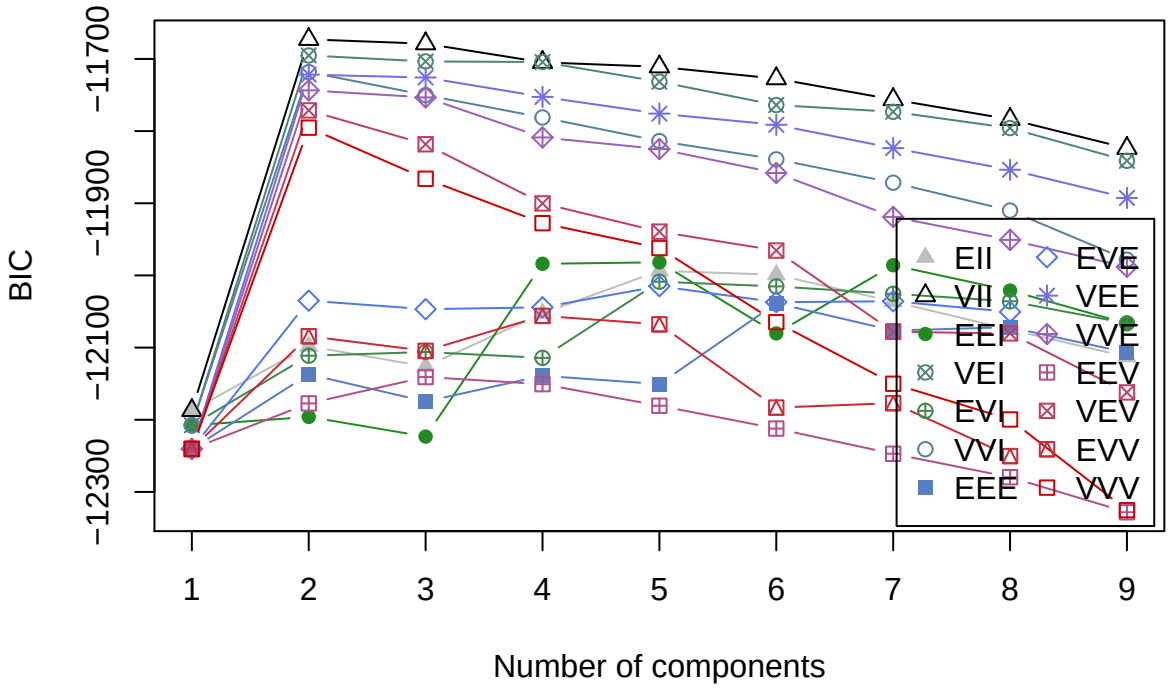
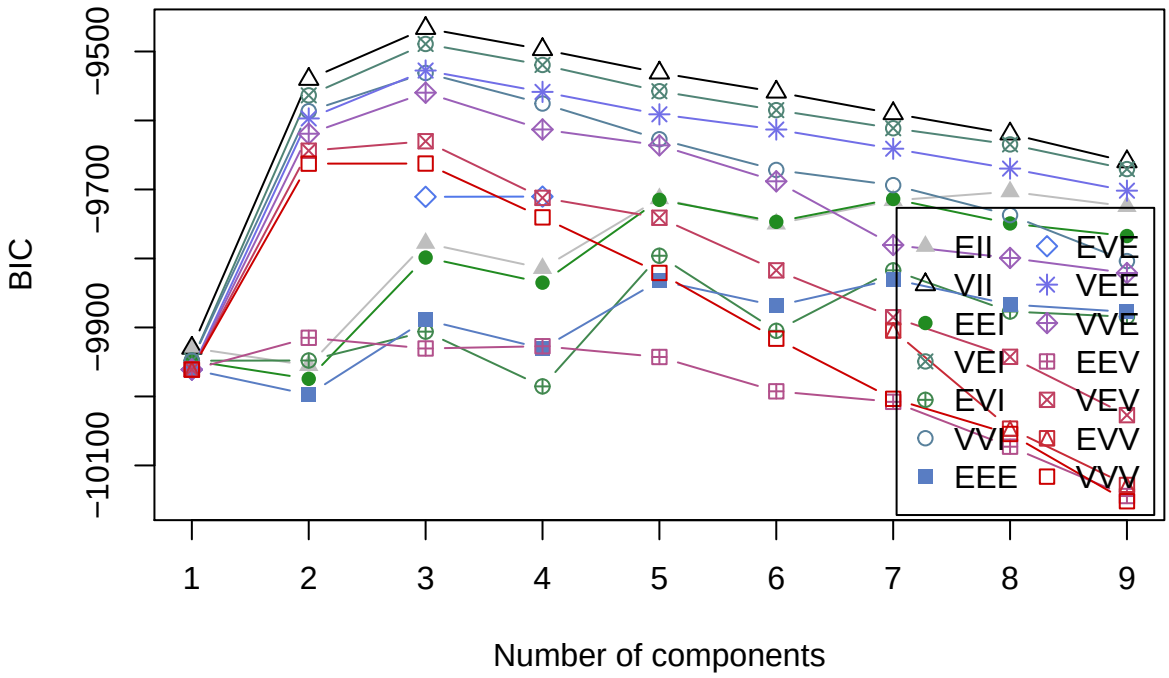


Table 6: Clustering Table Proportions for Down Class

Group	Proportion
1	0.483
2	0.491
3	0.026

Table 7: Model Metrics for Down Class

Metrics	Values
Model	VII
Log Like	-4672.34
n	441
df	5
BIC	-9466.46
ICL	-9646.07



According to the tables and graphs above, the *Up* class with the lowest BIC model has 2 groups with a spherical, varying volume model and a BIC of -11,672.89.

For the *Down* class, the lowest BIC model has 3 groups with a spherical, varying volume model and a BIC of -9,466.46.

This is displayed on the BIC plots with the VII model having the highest BIC for 2 and 3 components, respectfully.

Question 2(i) Part III

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

Answer

Calculated the TP, TN, FP, and FN based of the confusion matrix in the mclust model.It was just built on the training data.

Table 8: Metrics for Training Model

Metrics	Values
TPR	0.702
TNR	0.420
Training Error	0.424

Calculated the TP, TN, FP, and FN based of the confusion matrix in the mclust model. The summary of the model was updated with new data and class on the test data set.

Table 9: Metrics for Test Model

Metrics	Values
TPR	0.607
TNR	0.465

Metrics	Values
Test Error	0.452

Question 2(ii) Part I

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

Answer

Table 10: Summary of Mclust EDDA Model (Train)

	Down	Up
n	441	544
Proportion	0.45	0.55
Model	EII	EII
G	1	1

Table 11: Metric Summary of Mclust EDDA Model (Train)

Metrics	Values
Class Error	0.445
Brier Score	0.245
Log Like	-11026.144
BIC	-22128.107

Table 12: Weekly EDDA MclustDA Confusion Matrix

	Pred Down	Pred Up
Act Down	65	376
Act Up	62	482

Question 2(ii) Part II

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

Answer

According EDDA model above, is EII or spherical, equal volume with one group based on a BIC of -22,128.107.

Question 2(ii) Part III

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

Answer

I calculated the TP, TN, FP, and FN based of the confusion matrix in the EDDA mclust model. It was built on the training data.

Table 13: Metrics for Training EDDA Model

Metrics	Values
TPR	0.886
TNR	0.147
Training Error	0.445

Calculated the TP, TN, FP, and FN based of the confusion matrix in the EDDA mclust model. The summary of the model

was updated with new data and class on the test data set.

Table 14: Metrics for Test EDDA Model

Metrics	Values
TPR	0.607
TNR	0.465
Test Error	0.452

Question 2(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.*

Answer

Using formulas and code from Homework 4 with all Lag variables as predictors for each of the models, I summarized the Test Error, TPR, and TNR rates for each in a data frame.

Table 15: Model Summaries for Weekly Data All Lag Variables

Metrics	GLM	LDA	QDA	KNN	Mclust	Mclust_EDDA
Test Error	0.558	0.452	0.538	0.481	0.452	0.452
Test TPR	0.098	0.787	0.623	0.541	0.702	0.607
Test TNR	0.930	0.209	0.233	0.488	0.420	0.465

According to the summary above, LDA, Mclust, and Mclust_EDDA have the same and lowest test error rate, however the Mclust_EDDA have a more evenly distributed TPR and TNR than the other models. However, the Mclust model may be the best fit since the loss on the TNR rate comapred to the EDDA model isn't much, but the gain on the TPR rate is faily high.

Question 2(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.

Answer

Using the following variables to build the model based on Homework 4, which were chosen with all Lag variables, their interactions, squared and cubed. I used each of these variables as predictors in the new models and summazied the metrics in a data frame.

- Lag1
- Lag2
- Lag4
- Lag1 Squared
- Lag3 Squared
- Lag5 Squared
- Lag1:Lag3 Interaction
- Lag1:Lag3:Lag4 Interaction
- Lag2:Lag3:Lag4 Interaction
- Lag2:Lag3:Lag5 Interaction

I created a new data frame with the above variables and split the data into training and test data sets. These variables will be used for all the models in the previous question. In addition, I converted the *Direction* variable to to 1 for Up and 0 for Down for model simplicity.

Created the MclustDA model on the training data with the new variables using the same method as in the first model.

Table 16: Summary of Mclust Model 2 (Train)

	0	1
n	441	544
Proportion	0.45	0.55
Model	VVI	VVI
G	14	12

Table 17: Metric Summary of Mclust Model 2 (Train)

Metrics	Values
Class Error	0.328
Brier Score	0.253
Log Like	-22017.347
BIC	-47784.290

Table 18: Weekly MclustDA Confusion Matrix

	Pred Down	Pred Up
Act Down	284	157
Act Up	166	378

Summary of MclustDA Model shows that the model off the training data with the lowest BIC uses a diagonal, varying volume and shape model with 14 groups for the *Down* class and 12 for the *Up* class. The BIC is -47,784.29. This is lower than the previous model with just the Lag variables and less of a fit according to BIC, but better of a fit according to class error.

Created the MclustDA model setting the model type to 'EDDA' on the training data with the new variables using the same method as in the first model.

Table 19: Summary of Mclust EDDA Model 2 (Train)

	0	1
n	441	544
Proportion	0.45	0.55
Model	VVV	VVV
G	1	1

Table 20: Metric Summary of Mclust EDDA Model (Train)

Metrics	Values
Class Error	0.451
Brier Score	0.273
Log Like	-35773.906
BIC	-72443.856

Table 21: Weekly EDDA MclustDA Confusion Matrix

	Pred Down	Pred Up
Act Down	45	396
Act Up	48	496

Summary of EDDA MclustDA Model shows that the model off the training data with the lowest BIC uses a elliipsoidal, varying volume, shape, and orientation with 1 group for the *Down* and *Up* class. The BIC is -72,443.856. This is lower than the previous model with just the Lag variables and less of a fit according to BIC. It is also much worse than the Mclust model without the

EDDA model type.

I built all the models based on the methodology from the prior homework and built a data frame with a summary of the Test Error, Test TPR, and Test TNR and compared the results.

Table 22: Model Summaries for Weekly Data Select Lag Variables

Metrics	GLM	LDA	QDA	KNN	Mclust	Mclust_EDDA
Test Error	0.558	0.433	0.490	0.481	0.567	0.490
Test TPR	0.098	0.869	0.721	0.557	0.525	0.721
Test TNR	0.930	0.140	0.209	0.465	0.302	0.209

Comparing the 2 set of variables, the GLM model did not change. THe LDA model had a slightly better Test Error Rate, with a better TPR and worse TNR. QDA had a better Rest Error, with a better TPR and worse TNR. KNN Test Error stayed the same, with slightly better TPR and worse TNR. The Mclust DA model had a worse metrics for all 3 from the other model, while the EDDA model had a worse Test Error and TNR with a better TNR.

In addition, it looks like the Mclust model overfit the data as the traing error was lower than the first model, but the test error was higher than the first model.

Overall, similar to Homework 4, the KNN model seems to preform better with the second lowest Test Error, with a more balance TPR and TNR.

Source: An Introduction to Statistical Learning, MclustDA Part 1 Lecture, Dr. Saunders, Dakota State University, and Package 'mclust'