Modern Data Mining, HW 3

Jose Cervantez Bethany Hsaio Rob Kuan

11:59 pm, 03/17, 2024

Contents

1 PartI: Model Building

Multiple regression is one of the most popular methods used in statistics as well as in machine learning. We use linear models as a working model for its simplicity and interpretability. It is important that we use domain knowledge as much as we could to determine the form of the response as well as the function format for the factors. Then, when we have many possible features to be included in the working model it is inevitable that we need to choose a best possible model with a sensible criterion. Regularizations such as LASSO are introduced. Be aware that if a model selection is done formally or informally, the inferences obtained with the final `lm()` fit may not be valid. Some adjustment will be needed. This last step is beyond the scope of this class. Check the research line that Linda and collaborators have been working on.

The main job in this part is a rather involved case study about devastating covid19 pandemic. Please read through the case study first. This project is for sure a great one listed in your CV.

For covid case study, the major time and effort would be needed in EDA portion.

Answer:

See file covid_case_study_2024.pdf for our work and answers to Part I.

2 Part II: Logistic Regression

Logistic regression is used for modeling categorical response variables. The simplest scenario is how to identify risk factors of heart disease? In this case the response takes a possible value of `YES` or `NO`. Logit link function is used to connect the probability of one being a heart disease with other potential risk factors such as `blood pressure`, `cholestrol level`, `weight`. Maximum likelihood function is used to estimate unknown parameters. Inference is made based on the properties of MLE. We use AIC to help nailing down a useful final model. Predictions in categorical response case is also termed as `Classification` problems. One immediately application of logistic regression is to provide a simple yet powerful classification boundaries. Various metrics/criteria are proposed to evaluate the quality of a classification rule such as `False Positive`, `FDR` or `Mis-Classification Errors`.

LASSO with logistic regression is a powerful tool to get dimension reduction. We will not use it here in this work.

2.1 Framingham heart disease study

We will continue to use the Framingham Data (`Framingham.dat`) so that you are already familiar with the data and the variables. All the results are obtained through training data.

Liz is a patient with the following readings: `AGE=50, GENDER=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0`. We would be interested to predict Liz's outcome in heart disease.

To keep our answers consistent, use a subset of the data, and exclude anyone with a missing entry. For your convenience, we've loaded it here together with a brief summary about the data.

We note that this dataset contains 311 people diagnosed with heart disease and 1095 without heart disease.

```
   0    1
1095  311
```

After a quick cleaning up here is a summary about the data:

```
 HD             AGE            SEX            SBP            DBP
 0:1086   Min.   :45.0   FEMALE:730   Min.   : 90   Min.   : 50.0
 1: 307   1st Qu.:48.0   MALE  :663   1st Qu.:130   1st Qu.: 80.0
          Median :52.0                Median :142   Median : 90.0
          Mean   :52.4                Mean   :148   Mean   : 90.2
          3rd Qu.:56.0                3rd Qu.:160   3rd Qu.: 98.0
          Max.   :62.0                Max.   :300   Max.   :160.0
     CHOL           FRW            CIG
 Min.   : 96   Min.   : 52   Min.   : 0
 1st Qu.:200   1st Qu.: 94   1st Qu.: 0
 Median :230   Median :103   Median : 0
 Mean   :235   Mean   :105   Mean   : 8
 3rd Qu.:264   3rd Qu.:114   3rd Qu.:20
 Max.   :430   Max.   :222   Max.   :60
```

Lastly we would like to show five observations randomly chosen.

```
     HD AGE    SEX SBP DBP CHOL FRW CIG
643   1  61   MALE 140  68  248 104  20
11    0  45   MALE 110  88  183  90   0
576   1  58   MALE 150  95  296 100  15
560   1  59   MALE 260 130  246 111  20
702   0  45 FEMALE 122  74  178  88   5
```

2.1.1 Identify risk factors

2.1.1.1 Understand the likelihood function Conceptual questions to understand the building blocks of logistic regression. All the codes in this part should be hidden. We will use a small subset to run a logistic regression of `HD` vs. `SBP`.

i. Take a random subsample of size 5 from `hd_data_f` which only includes `HD` and `SBP`. Also set `set.seed(471)`. List the five observations neatly below. No code should be shown here.

Answer:

```
     HD SBP
643   1 140
11    0 110
576   1 150
560   1 260
702   0 122
```

---

ii. Write down the likelihood function using the five observations above.

Answer: The likelihood function using the five observations is:

$$\mathcal{L}\left(\beta_0, \beta_1 \mid \text{Data}\right) = \frac{e^{\beta_0+156\beta_1}}{1+e^{\beta_0+156\beta_1}} \cdot \frac{e^{\beta_0+164\beta_1}}{1+e^{\beta_0+164\beta_1}} \cdot \frac{1}{1+e^{\beta_0+156\beta_1}} \cdot \frac{1}{1+e^{\beta_0+138\beta_1}} \cdot \frac{1}{1+e^{\beta_0+155\beta_1}}$$

---

iii. Find the MLE based on this subset using glm(). Report the estimated logit function of `SBP` and the probability of `HD`=1. Briefly explain how the MLE are obtained based on ii. above.

Answer:

Table 1: Estimated Logit Function of SBP based on 5 observations

term	estimate	std.error	statistic	p.value
(Intercept)	-335	622609	-0.000538	0.9996
SBP	2.557	4730	0.0005405	0.9996

Based on maximum likelihood estimation, the equation for the estimated model is as follows.

$$Logit = -334.96 + 2.56 \times SBP$$

The probability of `HD`=1 is equal to:

$$\hat{P}(HD = 1 \mid SBP) = \frac{e^{-334.96+2.56\times SBP}}{1+e^{-334.96+2.56\times SBP}}$$

NOTE: the five samples that are randomly chosen, when classified, have **perfect separation**. This is why the parameters values are so extreme.

The maximum likelihood was obtained using the Fisher Scoring algorithm (which uses the Fisher Information Matrix) to estimate the parameters that are most likely to have generated the observed data. The algorithm iteratively updates the parameter estimates until the log-likelihood converges to a maximum value.

---

iv. Evaluate the probability of Liz having heart disease.

Answer:

The probability of Liz having heart disease is 2.22e-16 (which is an extremely small probability close to zero). Again, this is because the five data observations selected using seed(471) have **perfect separation** and can be classified with 100% accuracy.

```
## [1] "Liz's probability of heart disease is 2.22044604925031e-16"
```

2.1.1.2 Identify important risk factors for `Heart.Disease`. We focus on understanding the elements of basic inference method in this part. Let us start a fit with just one factor, `SBP`, and call it `fit1`. We then add one variable to this at a time from among the rest of the variables. For example

i. Which single variable would be the most important to add? Add it to your model, and call the new fit `fit2`.

We will pick up the variable either with highest $|z|$ value, or smallest p value. Report the summary of your `fit2` Note: One way to keep your output neat, we will suggest you using `xtable`. And here is the summary report looks like.

Answer:

SEX is the most important factor, as it is has the highest Z-value and results in the lowest AIC of all the other models. A summary of both `fit1` and `fit2` tables are below:

Table 2: Summary of `fit1` model

term	estimate	std.error	statistic	p.value
(Intercept)	-3.655	0.3479	-10.51	8.075e-26
SBP	0.01581	0.002222	7.118	1.097e-12

Table 3: Summary of `fit2` model

term	estimate	std.error	statistic	p.value
(Intercept)	-4.57	0.3897	-11.73	9.289e-32
SBP	0.01872	0.002324	8.053	8.071e-16
SEXMALE	0.9034	0.1398	6.464	1.02e-10

---

ii. Is the residual deviance of `fit2` always smaller than that of `fit1`? Why or why not?

Answer:

Yes, the residual deviance of `fit2` (no matter which predictor variables is used) will always be smaller than that of `fit1`. This is because adding an additional predictor will always add additional information to the model, which will reduce the residual deviance.

However, the AIC may not improve, because that takes into account the number of parameters in the model.

---

iii. Perform both the Wald test and the Likelihood ratio tests (Chi-Squared) to see if the added variable is significant at the .01 level. What are the p-values from each test? Are they the same?

Answer:

The p-values of both tests are displayed in the tables below. They are not the same, becuase the Wald test, the likelihood ratio test, and the score test (not mentioned in class) are all different ways of estimating the significance of the added variable. Even those these three tests are asympototically equivalent, in finite samples they may have different estimates.

Generally speaking, the likelihood ratio test is superior than the Wald Test because it does not require estimating an additional parameter (the standard deviation of the coefficient in `fit 2`).

Table 4: Likelihood Ratio Test (Chisquared)

#Df	LogLik	Df	Chisq	Pr(>Chisq)
3	-686.9	NA	NA	NA
2	-708.7	-1	43.7	3.828e-11

Table 5: Wald Test (Chisquared)

Res.Df	Df	Chisq	Pr(>Chisq)
1391	NA	NA	NA
1390	1	41.78	1.02e-10

2.1.1.3 Model building Start with all variables. Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).

i. Use backward selection method. Only keep variables whose coefficients are significantly different from 0 at .05 level. Kick out the variable with the largest p-value first, and then re-fit the model to see if there are other variables you want to kick out.

Answer:

Table 6: Summary of Best Model (Backwards Selection Method)

term	estimate	std.error	statistic	p.value
(Intercept)	-8.409	0.9086	-9.255	2.151e-20
AGE	0.05664	0.0145	3.906	9.377e-05
SEXMALE	0.9899	0.1451	6.824	8.842e-12
SBP	0.01696	0.002362	7.179	7.021e-13
CHOL	0.00448	0.001495	2.996	0.002737

---

ii. Use AIC as the criterion for model selection. Find a model with small AIC through exhaustive search. Does exhaustive search guarantee that the p-values for all the remaining variables are less than .05? Is our final model here the same as the model from backwards elimination?

Answer:

Exhaustive search does not gaurantee that all the remaining variables are less than 0.05. It only guarantees that the model with the smallest AIC is selected.

This model is different than the model selected from backwards elimination.

Table 7: Summary of Best Model (Exhaustive Search Method)

term	estimate	std.error	statistic	p.value
(Intercept)	-9.228	0.9962	-9.263	1.979e-20
AGE	0.06153	0.01478	4.164	3.122e-05
SEXMALE	0.9113	0.1571	5.8	6.633e-09
SBP	0.01597	0.002487	6.42	1.366e-10
CHOL	0.004493	0.001503	2.99	0.002794
FRW	0.006039	0.004004	1.508	0.1315
CIG	0.01228	0.006088	2.017	0.04369

---

iii. Use the model chosen from part ii. as the final model. Write a brief summary to describe important factors relating to Heart Diseases (i.e. the relationships between those variables in the model and heart disease). Give a definition of "important factors".

Answer:

Table 8: Increase in odds for each one unit increase in predictor

predictor	odds_increase
AGE	6.3%
SEXMALE	148.7%

predictor	odds_increase
SBP	1.6%
CHOL	0.5%
FRW	0.6%
CIG	1.2%

The important factors are not just coefficients that are statistically significant in the model. They must also have large betas that influence the probability of somebody having heart disease.

The final model shows that the following factors are significantly different than 0, and are correlated with increased odds of being diagnosed with heart disease: `SBP`, `SEX`, `AGE`, `CHOL`, and `CIG`.

- `SEX`: Being male vs. female increases your odds of heart disease by 148%
- `AGE`: For every year increase in age, the odds of having heart disease increase by 6.3%
- `SBP`: For every unit increase in systolic blood pressure, the odds of having heart disease increase by 1.6%
- `CHOL`: For every unit increase in cholestrol, the odds of having heart disease increase by 0.5%
- `CIG`: For every weekly self-reported cigarette smoked, the odds of having heart disease increase by 1.2%

Summary

When taking into account the beta coefficients, and the potential range of values for the individuals in the dataset, it seems that `SEX`, `AGE`, and `SBP` are the most important factors in predicting heart disease. `AGE` has a large absolute unit level impact on odds, and `SEX` and `AGE` also have a large impact when taking into account that they are continuous variables and there is a wide range of values in the dataset. `CIG` and `CHOL` may also be an important factor if individuals have extreme values for these variables.

Controlling for these other variables, we do not have enough evidence to reject the null hypothesis that the coefficient for `FRW` $= 0$, hence we cannot say if `FRW` is an important factor.

---

iv. What is the probability that Liz will have heart disease, according to our final model?

Answer:

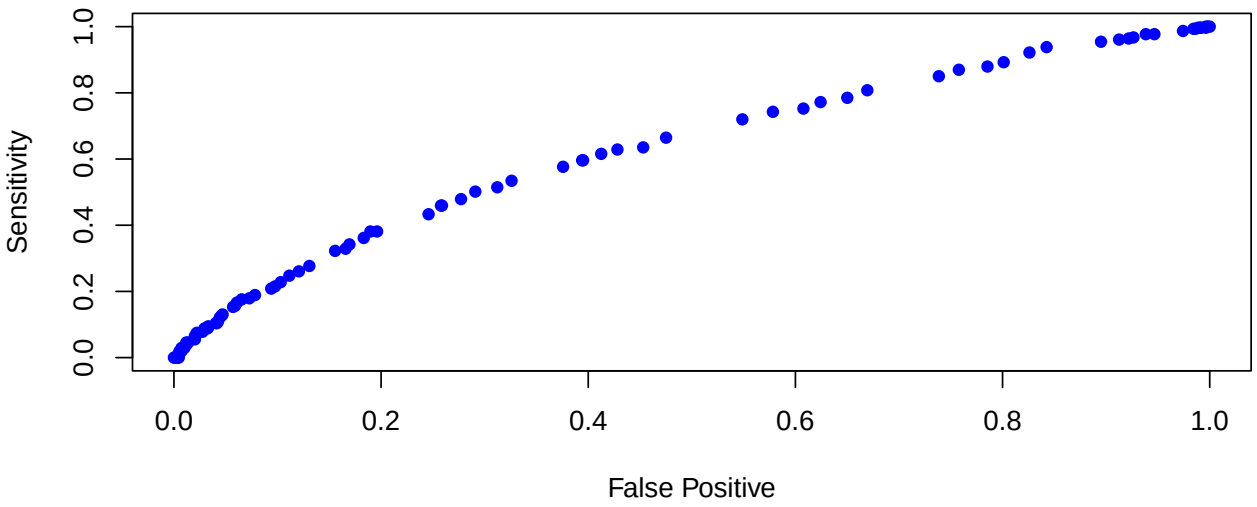
Liz's probability of heart disease is 3.46%.

2.1.2 Classification analysis

2.1.2.1 ROC/FDR

i. Display the ROC curve using `fit1`. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

Answer:




```
Optimal threshold: 0.298

True Positive Rate at this threshold: 0.215

False Positive Rate at this threshold: 0.0976
```

The ROC curve is a graphical representation of the true positive rate (TPR) and the false positive rate (FPR) for every possible threshold value. The ROC curve is also helpful for calculating metrics such as AUC (Area Under Curve) which can summarize the discrimination ability of the model.

The classifer where the false positive rate is less than 0.1 and the true positive rate is as high as possible is when the threshold is set at 0.298%:

$$\widehat{HD} = 1 \quad if \quad \hat{P}(HD = 1 \mid SBP) > 0.298$$

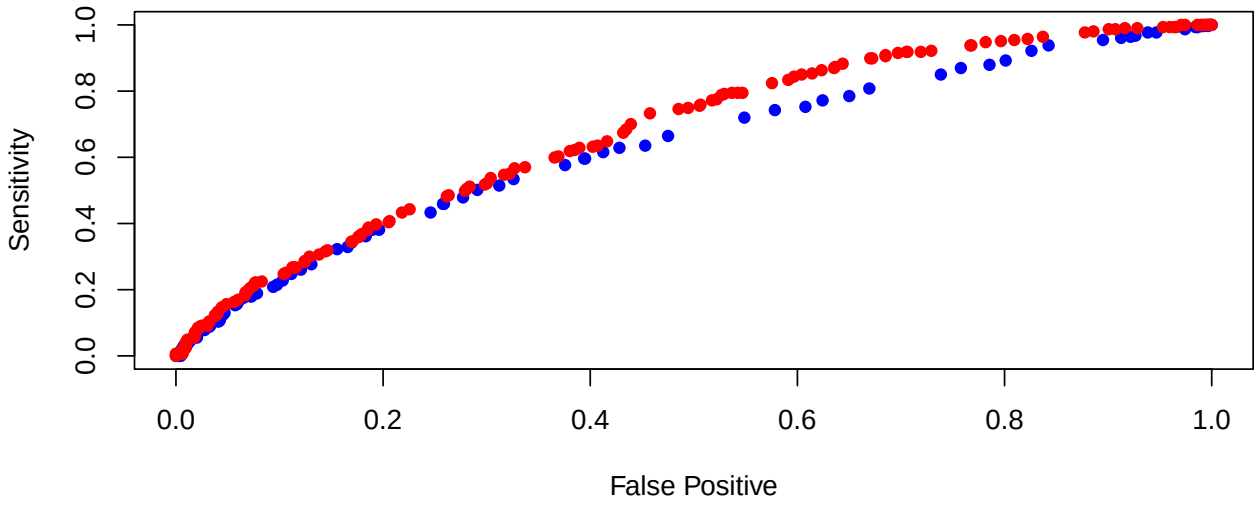
---

ii. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

Answer:

The ROC curve from `fit1` is always contained within the ROC curve from `fit2`. This is because `fit2` has all the predictor variables in `fit1`, and so the ROC curve from `fit2` will always have a higher true positive rate and a lower false positive rate than `fit1` because of the additional incremental information that can help improve the classifier.

See the chart below.



iii. Estimate the Positive Prediction Values and Negative Prediction Values for `fit1` and `fit2` using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values?

Answer:

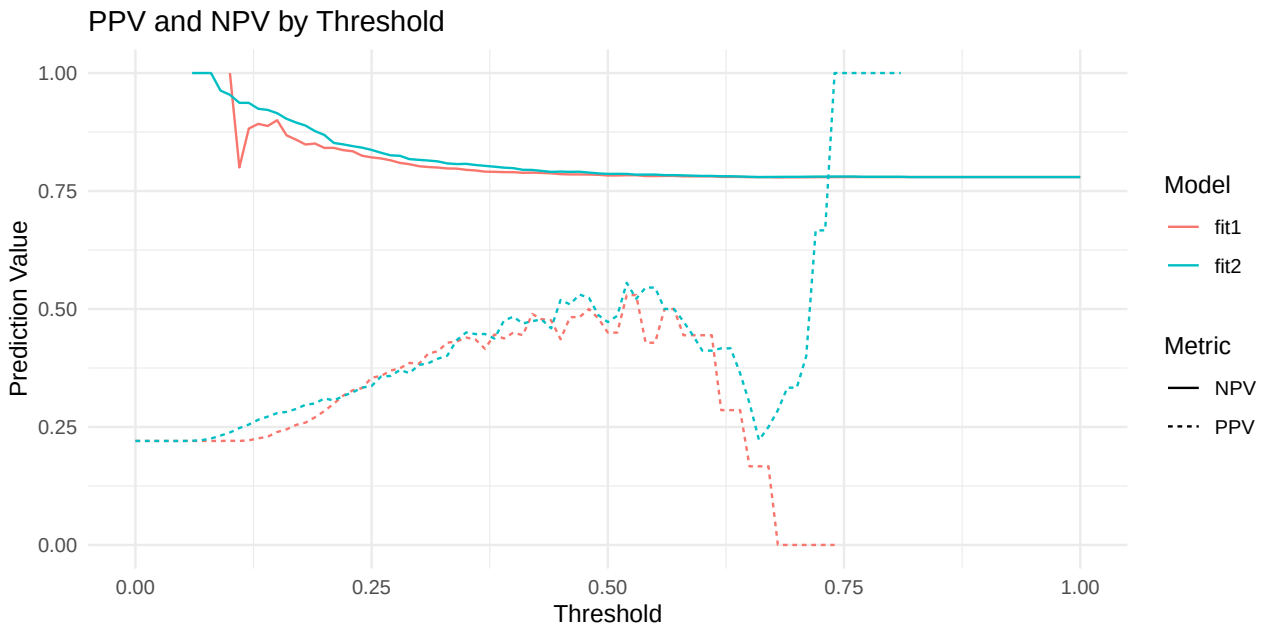
`fit2` is more desirable if we prioritize the Positive Prediction values. See table below:

Model	Positive_Prediction_Value	Negative_Prediction_Value
fit1	0.02932	0.9899
fit2	0.05537	0.9825

iv. For `fit1`: overlay two curves, but put the threshold over the probability function as the x-axis and positive prediction values and the negative prediction values as the y-axis. Overlay the same plot for `fit2`. Which model would you choose if the set of positive and negative prediction values are the concerns? If you can find an R package to do so, you may use it directly.

Answer:

If positive and negative prediction values are the concerns, I would choose `fit2` as it generally has higher positive prediction values (specifically for thresholds 0.1 to 0.5) and higher negative prediction values (on average) than `fit1`.


PPV and NPV by Threshold
Prediction Value
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Threshold
Model
fit1
fit2
Metric
NPV
PPV

2.1.2.2 Cost function/ Bayes Rule Bayes rules with risk ratio $\frac{a_{10}}{a_{01}} = 10$ or $\frac{a_{10}}{a_{01}} = 1$. Use your final model obtained from Part 1 to build a class of linear classifiers.

i. Write down the linear boundary for the Bayes classifier if the risk ratio of $a_{10}/a_{01} = 10$.

Answer:

If the risk ratio is $a_{10}/a_{01} = 10$, then we use caculate the optimal rule.

$$\begin{aligned}
\hat{Y} = 1 \quad \text{if} \quad & \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)} > \frac{a_{0,1}}{a_{1,0}} \\
\Leftrightarrow & P(Y=1 \mid X) > \frac{\frac{a_{0,1}}{a_{1,0}}}{1+\frac{a_{0,1}}{a_{1,0}}} \\
\Leftrightarrow & P(Y=1 \mid X) > \frac{\frac{1}{10}}{1+\frac{1}{10}} \\
\Leftrightarrow & P(Y=1 \mid X) > \frac{\frac{1}{10}}{1+\frac{1}{10}} \\
\Leftrightarrow & P(Y=1 \mid X) > \frac{1}{11}
\end{aligned}$$

Assuming that I am using the best fit model from Part 1, which is the model with the lowest AIC through exhaustive search, I can then establish the linear boundary below, expressed as variables of my final equation: (I've roudnded the coefficients below)

$$\text{logit} > \log\left(\frac{\frac{1}{11}}{\frac{10}{11}}\right) = -2.30$$

$$-9.23 + 0.06Age + 0.91SexMale + 0.016Sbp + 0.004Chol + 0.006Frw + 0.012Cig > -2.30$$

---

ii. What is your estimated weighted misclassification error for this given risk ratio?

Answer:

```
[1] "The estimated weighted misclassification error for the Bayes classifier is: 0.7143"
```

---

iii. How would you classify Liz under this classifier?

Answer:

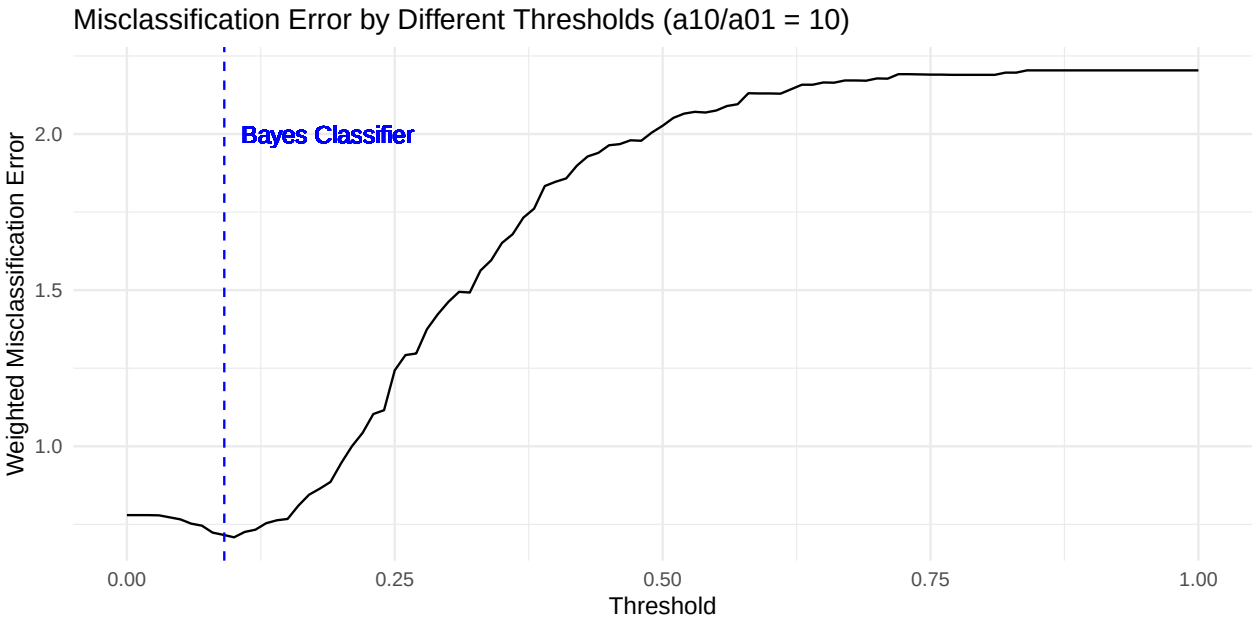
```
[1] "Liz's probability of heart disease is 0.0346"
[1] "Therefore, Liz would be classified as not having heart disease."
```

---

iv. Bayes rule gives us the best rule if we can estimate the probability of `HD-1` accurately. In practice we use logistic regression as our working model. How well does the Bayes rule work in practice? We hope to show in this example it works pretty well.

Now, draw two estimated curves where x = threshold, and y = misclassification errors, corresponding to the thresholding rule given in x-axis.

Answer:



---

v. Use weighted misclassification error, and set $a_{10}/a_{01} = 10$. How well does the Bayes rule classifier perform?

Answer:

The Bayes rule classifier performs vert well when the risk ratio is $a_{10}/a_{01} = 10$. The weighted misclassification error is the lowest when the threshold is set at 0.1 and minimum weighted misclassification error is equal to 0.709. The Bayes classifier results in a misclassification error of 0.714, which is very close to the minimum weighted misclassification error achievable in this dataset.

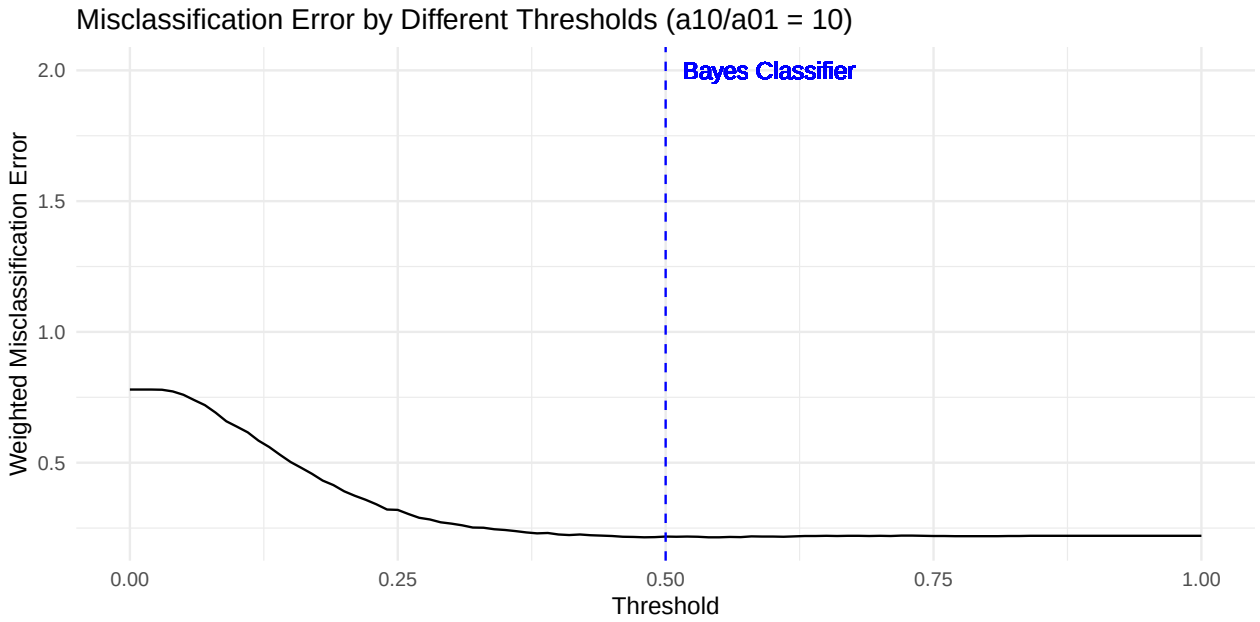
---

vi. Use weighted misclassification error, and set $a_{10}/a_{01} = 1$. How well does the Bayes rule classifier perform?

Answer:

The Bayes rule classifier performs also performs well when the risk ratio is $a_{10}/a_{01} = 1$. The Bayes classifier results in a misclassification error of 0.218, which is very close to the minimum weighted misclassification error achievable in this dataset (0.215).

See the graph below.


Misclassification Error by Different Thresholds (a10/a01 = 10)
Bayes Classifier
Weighted Misclassification Error
2.0
1.5
1.0
0.5
0.00
0.25
0.50
0.75
1.00
Threshold