Machine learning hw3 H24031354 統計 107 江冠駒

Goal: Use predictor with logistic regression to distinguish class “A” and class “B”.

1. Introduction to the dataset

The dataset we are going to analyze is an artificial well-known dataset. It contains some wave data. There are two classes with 21 variables and 100 noise variables. So we actually have 121 predictor variables and one response variable. Since our response has only two classes, we can say that this is a binary response dataset. It is a balanced classification dataset because the numbers of the subjects for both classes are almost the same. The number of observations is 33334. 10000 observations (about 30% of all data) will be the training data, 23334 observations (about 70% of all data) will be the test data. In training data, 5006 subjects are from class “A”, 4994 subjects are from class “B”. In testing data, 11760 subjects are from class”A”, 11574 subjects are from class”B”. No matter in training data or testing data, the ratio of training data size and testing data size is around 1:1, Almost the same ratio in our raw data. We have checked that no missing value in the dataset.

This how our data set looks like.

	variable									
observation		V1	V2	…	V21	alea1	alea2	…	alea100	classe	sample
	1	-0.38	-0.72		-0.69	0.06	0.32		0.07	A	learning
	2	-0.33	0.04		-0.10	0.36	0.74		0.71	A	learning
	3	0.48	0.94		0.98	0.19	0.25		0.21	A	learning
	4	0.57	1.15		-1.02	0.31	0.44		0.76	A	learning
	5	-1.99	-1.87		1.16	0.59	0.11		0.84	A	learning
	6	0.91	-1.18		-1.75	0.60	0.14		0.54	A	learning

V1,V2,…,V21 mean the predictor, alea1,alea2,…,alea100 mean the noise.
Classe is the response, its value is either “A” or “B”.
Sample indicates the observation belongs to training data or not, its value is either “learning” or “test”, “learning” means the observation will be split into training data, “test” means the observation will be split into testing data.

2. Data analysis

Because we have a lot of predictor variables, I assume that the 100 noise variables are not significant in classification, then I will run a two phase analysis. First, use all the predictor variables in logistic regression. After first stage, I use forward selection

to select important predictor variables, in this stage, I can also check the hypothesis is correct or not.

(1) logistic regression with all 121 predictor variables

Output of logistic regression:

```
Call:
glm(formula = classe ~ ., family = binomial, data = tr_mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5802  -0.1003  -0.0009   0.1798   3.1097

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.636983   0.833447  -6.763 1.35e-11 ***
v1           0.021127   0.043902   0.481 0.630357
v2          -0.046524   0.043962  -1.058 0.289923
v3          -0.066591   0.043102  -1.545 0.122356
v4          -0.294446   0.042262  -6.967 3.23e-12 ***
v5          -0.311524   0.041215  -7.558 4.08e-14 ***
v6          -0.034989   0.040739  -0.859 0.390426
v7           0.061602   0.041101   1.499 0.133924
v8           0.370755   0.043249   8.573  < 2e-16 ***
v9           0.599500   0.042986  13.946  < 2e-16 ***
v10          0.617721   0.042704  14.465  < 2e-16 ***
v11          0.852449   0.042300  20.153  < 2e-16 ***
v12          0.550198   0.042237  13.026  < 2e-16 ***
v13          0.200131   0.041597   4.811 1.50e-06 ***
v14         -0.154064   0.042339  -3.639 0.000274 ***
v15         -0.396582   0.042284  -9.379  < 2e-16 ***
v16         -0.496393   0.041910 -11.844  < 2e-16 ***
v17         -0.610071   0.043095 -14.157  < 2e-16 ***
v18         -0.455782   0.042695 -10.675  < 2e-16 ***
v19         -0.316138   0.044358  -7.127 1.03e-12 ***
v20         -0.152669   0.042404  -3.600 0.000318 ***
v21          0.063679   0.043529   1.463 0.143488
alea1        0.142670   0.149057   0.957 0.338491
alea2       -0.017536   0.149699  -0.117 0.906746
alea3        0.170593   0.151165   1.129 0.259100
alea4       -0.033302   0.151345  -0.220 0.825842
```

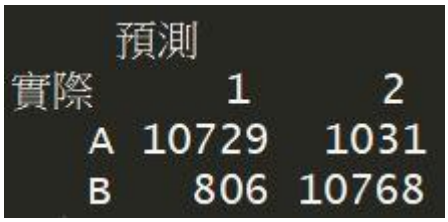
Although the R output is truncated, I can tell you that almost all the noise predictors are not significant(p-value>0.05) in this logistic regression model. In contrast, most original predictor variables(V1,V2,…,V21) are significant(p-value<0.05).

The AIC in this model:

```
    Null deviance: 13862.9  on 9999  degrees of freedom
Residual deviance:  3586.7  on 9878  degrees of freedom
AIC: 3830.7
```

We input our testing data into the logistic regression classifier we just trained and measure classification outcome.

Confusion matrix:



	預測	
實際	1	2
A	10729	1031
B	806	10768

There are 1837 misclassified examples among all 23,334 testing examples.

Accuracy	Error rate
0.9212737	7.87%

The accuracy is really high and error rate is pretty low, but maybe we can improve our model through feature selection to drop unimportant variables.

(2)Feature selection with forward selection

Output of forward selection:

```
Step:  AIC=3825.43
classe ~ v10 + v17 + v11 + v9 + v16 + v12 + v18 + v15 + v5 +
    v8 + v4 + v19 + v13 + v20 + v14
```

To minimize the BIC, forward selection only selects 15 predictor variables and there is no noise variables. So the hypothesis that noise variables are not important is correct.

Run logistic regression again with selected variables. Below is the new model output.

```
Call:
glm(formula = classe ~ v10 + v17 + v11 + v9 + v16 + v12 + v18 +
    v15 + v5 + v8 + v4 + v19 + v13 + v20 + v14, family = binomial,
    data = tr_mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4710  -0.1066  -0.0011   0.1913   3.0534

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.65949    0.34334 -16.484  < 2e-16 ***
v10          0.60737    0.04135  14.689  < 2e-16 ***
v17         -0.59536    0.04181 -14.239  < 2e-16 ***
v11          0.83929    0.04066  20.641  < 2e-16 ***
v9           0.58022    0.04136  14.027  < 2e-16 ***
v16         -0.48211    0.04048 -11.910  < 2e-16 ***
v12          0.53901    0.04064  13.262  < 2e-16 ***
v18         -0.44028    0.04161 -10.581  < 2e-16 ***
v15         -0.38726    0.04086  -9.478  < 2e-16 ***
v5          -0.29295    0.03913  -7.487 7.02e-14 ***
v8           0.36292    0.04161   8.721  < 2e-16 ***
v4          -0.29352    0.04083  -7.188 6.56e-13 ***
v19         -0.30429    0.04304  -7.070 1.55e-12 ***
v13          0.18547    0.03971   4.670 3.01e-06 ***
v20         -0.14752    0.04146  -3.558 0.000374 ***
v14         -0.13943    0.04016  -3.472 0.000517 ***
```

We can see that all the predictor variables are significant(p-value <0.05)

The AIC in this model:

```
    Null deviance: 13862.9  on 9999  degrees of freedom
Residual deviance:  3678.1  on 9984  degrees of freedom
AIC: 3710.1
```

Compare to the original model, the AIC have improved from 3830.7 to 3710.1, our new model is better.

We input our testing data into the new logistic regression classifier we just trained and measure classification outcome.

Confusion matrix:

There are 1788 misclassified examples among all 23,334 testing examples, better than the original model.

Accuracy	Error rate
0.9233736	7.66%

Both accuracy and error rate improve compared to the original model.

3. Conclusion

Feature selection can help us to train a better classifier, thus we can get our prediction outcome better.