Biostat626 Midterm1 Writeup Submit

Mingrui Li

4/14/2023

1 Task 1

1. Lasso was used to select variables. From the plot, we noticed that the misclassification error is close to 0, which means the accuracy is pretty high for binary classifier.

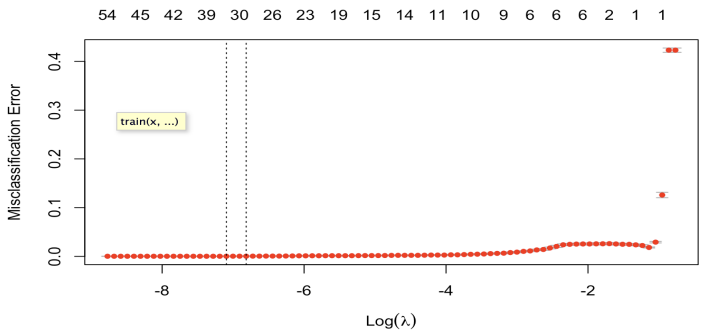

Figure 1: Task1 lasso

2. The baseline algorithms used for binary classifier are random forests, XGBoost, SVM and Neural Network. The table shows that SVM and Neural Network perform the best with accuracy 1 while other algorithms also have accuracy close to 1.

A tibble: 5 × 2

Algorithm <chr>	CV Accuracy <dbl>
Lasso	0.9997425
Random Forest	0.9996139
XGBoost	0.9994850
SVM	1.0000000
Neural Network	1.0000000

5 rows

Figure 2: Task1 table

3. The final algorithm used for binary classifier is chosen to be SVM with radial kernel.

2 Task 2

1. Lasso and PCA are used to select variables. PCA performed faster than lasso. Thus, PCA with a threshold 0.9 ->0.8 was used.
2. The baseline algorithms used for multi-class classifier are random forests, XGBoost, SVM and Neural Network. The table shows that SVM performs the best with accuracy 0.9289813.

A tibble: 5 × 2

Algorithm <chr>	CV Accuracy <dbl>
Lasso	0.9997425
Random Forest	0.9120122
XGBoost	0.9125013
SVM	0.9289813
Neural Network	0.9209472

5 rows

Figure 3: Task2 table

3. The final algorithm used for multi-class classifier is SVM with radial kernel.

3 Leaderboard performance

SVM with PCA threshold 0.9	SVM with PCA threshold 0.8	SVM with PCA threshold 0.8 with repeated cv
0.910	0.880	0.873

Figure 4: Leaderboard performance for task2

For task2, I first select variables using PCA with threshold 0.910, it gave an accuracy of 0.910. Then I tried threshold 0.8 which gave a lower accuracy of 0.880. The low accuracy might due to overfitting, thus, repeated cv was used with PCA threshold 0.8 to improve fitness. However, this gave a lower accuracy of 0.873, which shows that selecting variables may not work well for this classification.

4 Discussion

The final result for binary classifier is 0.999 and is 0.873 for multi-class classifier. Selecting variables may not fit this classification. One way to improve the classification accuracy is not

to select variables to fit models. Further I would try to split the ordinary training data into 70%:30% and 80%:20% to see if this would give a better accuracy.

5 URL

https://github.com/imingrui/biostat626midterm1