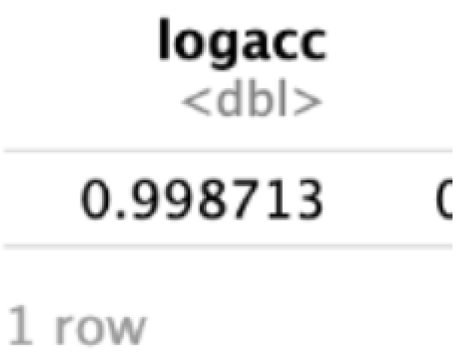
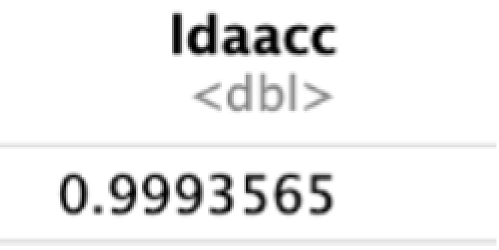
Task 1

Baseline algorithms

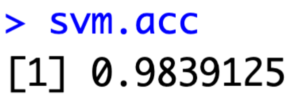
For task 1, I used the logistic regression model, SVM, and LDA as my baseline models. It turns out that these models have very high accuracy (around 0.99) on training data even without feature selection, which might be due to overfitting.



```
logacc
<dbl>
0.998713
1 row
```



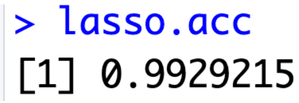
```
ldaacc
<dbl>
0.9993565
```



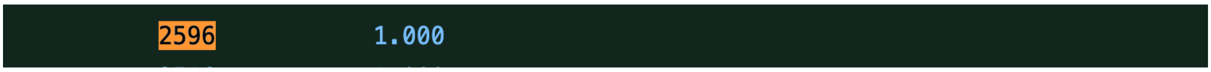
```
> svm.acc
[1] 0.9839125
```

Final algorithm

In order to do the feature selection, I applied the L1 regularized logistic model, which includes a penalized term. After conducting 5 fold cross validation on this model, it still achieved high accuracy (0.99) on training data, and it also performed very well on test data (100%).



```
> lasso.acc
[1] 0.9929215
```

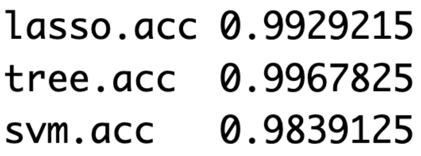


```
        2596          1.000
```

Task 2

Baseline algorithms

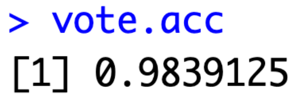
For task 2, I added random forest together with L1 regularized multinomial logistic model and SVM. After 5 fold cross validation, These three models still achieved very high accuracy on training data. Although the randomforest achieved the highest accuracy, I still chose the L1 logistic model since the tree method may have overfitting issues from handling noises in the dataset



```
lasso.acc 0.9929215
tree.acc  0.9967825
svm.acc   0.9839125
```

Final algorithm

The L1 logistic model achieved 0.955 accuracy on the first submission. In order to improve performance, I tried to use the gini importance of random forest to manually select most “important” features. After extracting these features with gini importance greater than 10, I fitted the new L1 model, random forest model and SVM model with them and implemented a community voting scheme, which combined the prediction results of three models together and “voted” for the final label (with mode). However, this voting model didn’t improve the prediction results, so I kept with the L1 logistic model.



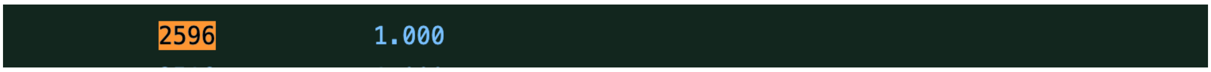
```
> vote.acc
[1] 0.9839125
```

Leaderboard Performance

For the binary classification, I achieved 100% accuracy on the first submission, so I didn’t continue with improvement.

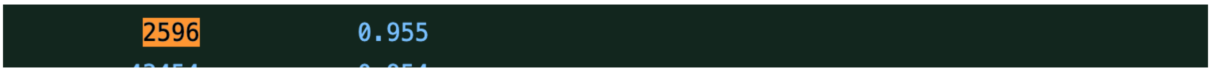
In the multiclass classification, I got 0.955 accuracy on the first submission. Then, I tried to implement a community voting scheme, but it didn’t improve the final result, so I chose the L1 logistic model as my final algorithm.

Task 1:

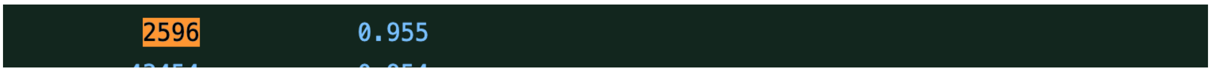


```
2596          1.000
```

Task 2:



```
2596          0.955
```



```
2596          0.955
```

Future Plans

The final result of task 1 is satisfactory since it gets a 100% accuracy. However, the result of task 2 is quite confusing. Although I tried different methods to improve, the final results didn’t make much changes. To further improve the accuracy on the test data, I’m thinking of using different feature selection methods such as wrapper and embedded methods besides using L1 regularization to reduce the overfitting problem (high accuracy on training data but relatively low accuracy on test data). In addition, I didn’t try other algorithms such as neural network and adaboost, which may further improve classification performance. I will try to use them in the future.