Machine Learning Assignment - 2

Siddharth Reddy Aleti

20010112-T150

sial21@student.bth.se

Akhilesh Yadlapalli

20010311-T050

akya21@student.bth.se

I. Introduction

In this assignment the task is to compare and find if there is any computational and predictive performance difference between the selected three supervised machine learning algorithms by detecting the spam mails on the given spam base dataset. We have selected **Random Forest**, **Support Vector Machine** and **Logistic Regression** algorithms to perform the given task. We ran a stratified ten-fold cross-validation tests to find the Training time, Accuracy and F-measure of each selected machine learning algorithms and presented them in different tables as mentioned in the table in example 12.4 of the main literature [1]. Then with Freidman and Nemenyi tests we have to check whether the above algorithms perform same or different.

II. Concepts and Hypothesis

A. *Performing a stratified ten-fold cross-validation tests.*

For performing a stratified ten-fold cross-validation tests, we imported StratifiedKFold class from module sklearn.model_selection, which is a type of cross-validation that ensures that each fold contains a representative proportion of the different classes in the data set.

We initialized a cross-validation object called 'skf' using stratified ten-fold cross-validation with parameters n_splits=10. This means that the data is split into ten folds, with each fold containing roughly the same proportions of the different classes in the target variable.

The code then enters a loop over the folds of the cross-validation object. For each fold, it divides the data into training and test sets using the indices provided by the cross-validation object.

B. *Friedman Test.*

This test works on the idea of ranking the performances of k algorithms per data set, from best performance (rank 1) to worst performance (rank k) [1] with respective average ranks. Then we have to state the **hypothesis** as null and an alternate hypothesis. In this task the null hypothesis h_0 is that all the algorithms work with same performance. The alternate hypothesis h_1 is the opposite to null that is, the algorithms have different performances.

To test the null hypothesis, we need to calculate the following formulas [1],

1. Average rank $\bar{R} = \frac{1}{nk}\sum_{ij} R_{ij} = \frac{k+1}{2}$
2. Avg ranks sum of squared differences = $n\sum_j \left(R_j - \bar{R}\right)^2$
3. All ranks sum of squared differences

$= \frac{1}{n(k-1)}\sum_{ij}\left(R_{ij} - \bar{R}\right)^2$

Now we need to find the Friedman statistic, it is the ratio between the second and third formula. Then we find the critical value with the values of k, n and α. Where k=3, n=10 and $\alpha = 0.05$ for this task. So according to this the critical value level is 7.8 [1]. If the Friedman statistic is greater than the critical value then we can reject the null hypothesis and say that the algorithms work with different computational and predictive performance. To find which algorithm is the better one we need to apply a post-hoc test that is Nemenyi test because Friedman test only tells us that there is a difference in performance among the algorithms.

C. *Nemenyi Test.*

This test is a post-hoc test to Friedman test and works on the idea to calculate the Critical Difference (CD) and compare it against the difference of the average ranks of two algorithms [1]. This test is based on pairwise performance.

- Critical Difference (CD) = $q_\alpha \sqrt{\frac{k(k+1)}{6n}}$

Where q_α is depended on both α and k values. The Critical Difference is compared with every unique pair of algorithms "average-rank ($\bar{R}$) difference". If it is greater than the Critical Difference then in that pair, the algorithm with highest average-rank ($\bar{R}$) is the better algorithm than other algorithms. As our task includes α = 0.05and k = 3, q_α will be 2.343 and our Critical Difference (CD) is 1.047.

III. RESULTS

A. Training Time.

The training time results obtained after conducting stratified ten-fold cross-validation tests for the three algorithms can be seen in the below table-1,

Training Times:

	Random Forest	Logistic Regression	SVC
Fold 1	0.726730	0.068111	0.462447
Fold 2	0.724722	0.072702	0.463061
Fold 3	0.807784	0.078074	0.471758
Fold 4	0.758875	0.087958	0.468581
Fold 5	0.735367	0.079899	0.467640
Fold 6	0.717970	0.084311	0.460372
Fold 7	0.767467	0.084429	0.467851
Fold 8	0.727168	0.078673	0.474165
Fold 9	0.767333	0.076482	0.467098
Fold 10	0.726406	0.132684	0.468993
Average	0.745982	0.084332	0.467197
Standard Deviation	0.027130	0.017035	0.004017

Table-1

For this task we took α=0.05, k=3 and n=10, the critical value is 7.8. After ranking the above table according to Friedman test, we can obtain the average rank for each algorithm. So based on the Friedman statistic formula we get the statistic value as 20.

As the Friedman statistic is greater than the critical value, we can reject the null hypothesis and say that the algorithms does not perform same. Now we can conduct post-hoc test pairwise Nemenyi test to check for a significant difference between the algorithms.

As we got the Critical Difference as 1.047, the difference between the average ranks of Random Forest and Logistic Regression is 2. This pair's difference is the only one greater than the CD. So, we can say that there is a significant performance difference between the Random Forest and Logistic Regression algorithms in terms of training time.

----------training times----------

	Random Forest	Logistic Regression	SVC
0	0.726730 (3)	0.068111 (1)	0.462447 (2)
1	0.724722 (3)	0.072702 (1)	0.463061 (2)
2	0.807784 (3)	0.078074 (1)	0.471758 (2)
3	0.758875 (3)	0.087958 (1)	0.468581 (2)
4	0.735367 (3)	0.079899 (1)	0.467640 (2)
5	0.717970 (3)	0.084311 (1)	0.460372 (2)
6	0.767467 (3)	0.084429 (1)	0.467851 (2)
7	0.727168 (3)	0.078673 (1)	0.474165 (2)
8	0.767333 (3)	0.076482 (1)	0.467098 (2)
9	0.726406 (3)	0.132684 (1)	0.468993 (2)
10	3.0	1.0	2.0

friedman statistic : 20.0
The critical value for k = 3 and n = 10 at the alpha = 0.05 level is 7.8 i.e 20.0 > 7.8
The null hypothesis is rejected that is all Algorithms doesnot perform equally
The critical difference is : 1.0478214542564015
Random Forest and Logistic Regression do not perform equally

Table-2

B. Accuracy.

The Accuracy results obtained after conducting stratified ten-fold cross-validation tests for the three algorithms on spam base dataset can be seen in the table-3.

Accuracies:

	Random Forest	Logistic Regression	SVC
Fold 1	0.954447	0.913232	0.919740
Fold 2	0.954348	0.923913	0.928261
Fold 3	0.960870	0.928261	0.939130
Fold 4	0.963043	0.932609	0.921739
Fold 5	0.945652	0.908696	0.939130
Fold 6	0.965217	0.936957	0.947826
Fold 7	0.950000	0.934783	0.926087
Fold 8	0.954348	0.926087	0.936957
Fold 9	0.945652	0.939130	0.945652
Fold 10	0.950000	0.910870	0.934783
Average	0.954358	0.925454	0.933930
Standard Deviation	0.006522	0.010541	0.009155

Table-3

Based on table-4, we get the Friedman statistic as 14.84 which is greater than the critical value 7.8. So, we reject the null hypothesis and conduct Nemenyi test.

The average rank difference of both Random Forest-Logistic Regression and Random Forest-Support Vector Machine is greater than the CD. So, the above pairs can be said to be significantly different performing pairs of algorithms.

----------Accuracies----------

	Random Forest	Logistic Regression	SVC
0	0.954447 (1)	0.913232 (3)	0.919740 (2)
1	0.954348 (1)	0.923913 (3)	0.928261 (2)
2	0.960870 (1)	0.928261 (3)	0.939130 (2)
3	0.963043 (1)	0.932609 (2)	0.921739 (3)
4	0.945652 (1)	0.908696 (3)	0.939130 (2)
5	0.965217 (1)	0.936957 (3)	0.947826 (2)
6	0.950000 (1)	0.934783 (2)	0.926087 (3)
7	0.954348 (1)	0.926087 (3)	0.936957 (2)
8	0.945652 (2)	0.939130 (3)	0.945652 (2)
9	0.950000 (1)	0.910870 (3)	0.934783 (2)
10	1.1	2.8	2.2

friedman statistic : 14.841930116472545
The critical value for k = 3 and n = 10 at the alpha = 0.05 level is 7.8 i.e 14.841930116472545 > 7.8
The null hypothesis is rejected that is all Algorithms doesnot perform equally
The critical difference is : 1.0478214542564015
Random Forest and Logistic Regression do not perform equally
SVC and Random Forest do not perform equally

Table-4

C. F-Measure.

The F-measure results obtained after conducting stratified ten-fold cross-validation tests for the three algorithms on spam base dataset can be seen in the below table-5,

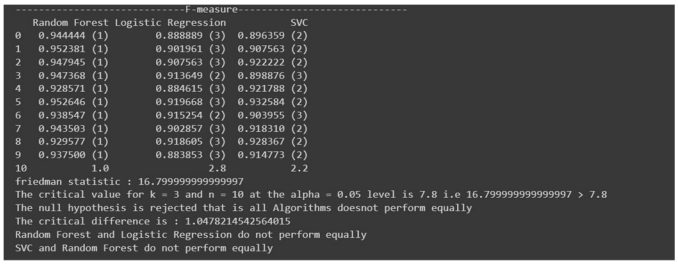
F-Measures:

	Random Forest	Logistic Regression	SVC
Fold 1	0.941504	0.888889	0.896359
Fold 2	0.940510	0.901961	0.907563
Fold 3	0.950000	0.907563	0.922222
Fold 4	0.952646	0.913649	0.898876
Fold 5	0.931129	0.884615	0.921788
Fold 6	0.955307	0.919668	0.932584
Fold 7	0.935211	0.915254	0.903955
Fold 8	0.941176	0.902857	0.918310
Fold 9	0.929178	0.918605	0.928367
Fold 10	0.935211	0.883853	0.914773
Average	0.941187	0.903691	0.914480
Standard Deviation	0.008519	0.013060	0.011744

Table-5

From Table-6 we can say that that Friedman statistic is 16.79 is greater than critical value of 7.8. So, we reject the null hypothesis and conduct Nemenyi test.

The average rank difference of both Random Forest-Logistic Regression and Random Forest-Support Vector Machine is greater than the CD. So, the above pairs can be said to be significantly different performing pairs of algorithms.



```
----------------------------F-measure----------------------------
   Random Forest Logistic Regression          SVC
0   0.944444 (1)        0.888889 (3)  0.896359 (2)
1   0.952381 (1)        0.901961 (3)  0.907563 (2)
2   0.947945 (1)        0.907563 (3)  0.922222 (2)
3   0.947368 (1)        0.913649 (2)  0.898876 (3)
4   0.928571 (1)        0.884615 (3)  0.921788 (2)
5   0.952646 (1)        0.919668 (3)  0.932584 (2)
6   0.938547 (1)        0.915254 (2)  0.903955 (3)
7   0.943503 (1)        0.902857 (3)  0.918310 (2)
8   0.929577 (1)        0.918605 (3)  0.928367 (2)
9   0.937500 (1)        0.883853 (3)  0.914773 (2)
10         1.0                 2.8           2.2
friedman statistic : 16.799999999999997
The critical value for k = 3 and n = 10 at the alpha = 0.05 level is 7.8 i.e 16.799999999999997 > 7.8
The null hypothesis is rejected that is all Algorithms doesnot perform equally
The critical difference is : 1.0478214542564015
Random Forest and Logistic Regression do not perform equally
SVC and Random Forest do not perform equally
```

Table-6

IV. Conclusion

Based on the results, for Training time only the Random Forest and Logistic Regression algorithms are having significantly different performance and Logistic Regression has the better training time in the above pair.

For Accuracy and F-measure, both Random Forest-Logistic Regression and Random Forest-Support Vector Machine pairs are having a significantly different performance based on the tests conducted and Random Forest algorithm has the better Accuracy and F-measure in the above pairs of algorithms.

References

[1] Flach, Peter. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press, 2012.