Assignment 2 Report

DV2578-Machine Learning

Rong Peng
19951120-7004
Blekinge Institute of Technology

I. Introduction

In this experiment, used CART (Classification And Regression Tree), LDA(Linear Discriminant Analysis) and LR(Linear Regression) three supervised classification learning algorithms on the Spam detection task. And conducted the Friedman test to determine whether the average ranks as a whole display significant differences on the 0.05 alpha level, and if so use Nemenyi test to calculate critical difference in order to determine which algorithms perform significantly different from each other. Lastly, compared their computational and predictive performance. Program Language chose Python, related libraries chose Sklearn, Pandas and Numpy.

II. Experiment Implementation

A. The Stratified ten-fold Cross-validation Tests Results

Firstly, I used Pandas built-in function *read_csv* to load the Spambase dataset[1], then make an array slice to separate the {1,0} classification results as Y and the rest data as X. Secondly, set a number for random seed, in this experiment set seed=7, and then used sklearn built-in function *train_test_split* to separate dataset as X_train, X_validation, Y_train, Y_validation. Thirdly, used sklearn built-in function *KFold*, *cross_val _score* to run ten-fold cross-validation repeat for each model.
The following table gives a possible result of evaluating three learning algorithms on a data with ten-fold cross-validation:

Cross-validation

Fold	Linear Discriminant Analysis	CART	Linear Regression
1	0.8804	0.9239	0.9157
2	0.9076	0.9130	0.9375
3	0.8668	0.9158	0.913
4	0.9049	0.9076	0.9239
5	0.9239	0.9375	0.9402
6	0.8696	0.8967	0.9266
7	0.8723	0.9158	0.9185
8	0.8995	0.9429	0.9484
9	0.8913	0.9158	0.9212
10	0.8614	0.9022	0.9130
avg	0.8878	0.9171	0.9258
stdev	0.0197	0.0137	0.0117

The last two lines show the average and standard deviation over all ten folds. The standard deviation reflects the degree of dispersion of the data set under three models .The LR achieves the best result, LDA presents the worst.

B. Friedman Test Results

Used the data from the *Cross-validation* table, assuming it comes from different data sets, and the following table shows the ranks in brackets:

Friedman test

Data set	Linear Discriminant Analysis	CART	Linear Regression
1	0.8804(3)	0.9239(1)	0.9157(2)
2	0.9076(3)	0.9130(2)	0.9375(1)
3	0.8668(3)	0.9158(1)	0.9130(2)
4	0.9049(3)	0.9076(2)	0.9239(1)
5	0.9239(3)	0.9375(2)	0.9402(1)
6	0.8696(3)	0.8967(2)	0.9266(1)
7	0.8723(3)	0.9158(2)	0.9185(1)
8	0.8995(3)	0.9429(2)	0.9484(1)
9	0.8913(3)	0.9158(2)	0.9212(1)
10	0.8614(3)	0.9022(2)	0.9130(1)
avg rank	3	1.8	1.2

$\bar{R}$=2 , $^{n\sum_j (R_j-\bar{R})^2}$ = 16.8 and $\frac{1}{n(k-1)}\sum_{ij}(R_{ij}-\bar{R})^2$=1, so Friedman statistic is 16.8. The critical value for k=3 and n=10 at 0.05 alpha level is 7.8, therefore we can reject the null hypothesis that all algorithms perform equally.

C. Nemenyi Tests Results

The Nemenyi test format as follows[2]:

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

In this experiment, qα=0.05,k=3,n=10,so CD=1.047, which means the average ranks 3, 1.8, 1.2, the differences between the LDA and CART or the LDA and LR exceeds the critical difference. So LDA performs significantly different the rest two algorithms.

D. Computational And Predictive Performance Result

Used the function *time.clock* to calculate the start and end time to get the models training time, and repeat in five time to get the average training time. Used scikit-learning built-in functions to train models, generate classification reports and get accuracy scores respectively. The results are shown in the following table:

Accuracy&F1-score

Models	Accuracy	F1-score
LDA	0.8719	0.87
CART	0.9186	0.91
LR	0.9164	0.92

Trainning Time

Time(sec)	LDA	CART	LR
1	0.0148	0.0689	0.0913
2	0.0145	0.0749	0.0878
3	0.0146	0.0722	0.0905
4	0.0143	0.0669	0.0868
5	0.0147	0.0667	0.0925
avg	0.0146	0.0699	0.0898

III. Discussion

This experiment only tried three supervised classification algorithms. In the future work, can try some other supervised algorithms such as SVM, KNN, Naïve Bayes, compare their significance difference with LR and CART.

IV. Conclusion

From the stratified ten-fold cross-validation tests results finds that LDA achieves the highest standard deviation, presents the worst, and LR achieves the best. However we need to use the significance testing to continue verifying. Friedman test is a suitable test, it tells us the significant differences. From the calculation result 16.8 is greater than 7.8, which means three algorithms perform differently. Therefore further analysis is needed on a pairwise level. Nemenyi test can compare the critical difference between two algorithms. The LDA's difference between LR is 1.8, between CART is 1.2. Clearly LDA exceeds the calculated critical difference(CD=1.047), which means LDA performing significantly different from the rest two algorithms.
From the computational and predictive performance results, can know that CART's accuracy achieves the highest score, which means it is good at predicting samples are correctly predicted. F1-Score is a comprehensive standard, obviously, LR achieves the best. From the training time table, LR requires the most time, CART is faster than LR, LDA requires the minimum time.

REFERENCES

[1] "UCI Machine Learning Repository: Spambase Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/spambase. [Accessed: 20-Nov-2018].

[2] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012. Page 350-357