DV2578-Assignment 2

Patamsetti Naga Vishnu Chakravarthi *Department of Computer Science Blekinge Institute of Technology*
Karlskrona, Sweden. napa18@student.bth.se

I. Introduction

The main aim of this assignment is done in python. The considered algorithms are State Vector Machine, Logistic regression and KNN. These are used to produce the results. From the above algorithms calculating and comparing of predictive performance, F-measure, computational time are done. Therefore, results will be reported by performance of three tests.

1. Friedman test
2. Stratified 10-fold cross validation test and
3. Average rank with a significant difference of the alpha value being 0.05.

The mentioned three algorithms State Vector Machine, logistic regression and KNN are commonly used in Machine Learning. By comparing the three algorithms each varies from others State Vector Machine works is not a appropriate algorithm for large amount of data but acts efficiently with high dimensional data. KNN is a simple algorithm and computationally expensive. Logistic regression is much faster in training and giving out results. As the difference is observed in three algorithm behavior, I had chosen them.

From the above three algorithms Accuracy, Training time and F-measure can be calculated. By using F-measure precision and recall can be calculated.

II. Methodology

1) At first required data and the libraries are imported. This is done by using built-in python library.
2) By using of stratified- fold library the considered data will be divided in to 10 folds. Each of the fold has 460 instances.
3) From among the 10 folds, 9-folds are for training and the remaining 1- fold is used for testing.
4) The previously done process will be repeated for a couple of times (10). So, at each time the test fold will get changed. And the whole process in the training data will be performed on three algorithms.
5) Training time will be calculated after predicting the data for State Vector Machine. This will be done by the using of module import time.
6) In a same way, this process will also be done for KNN and Logistic regression. This is done to find the recall, precision accuracy, f-measure confusion matrix is used.
7) Results will be obtained from the performance test by conducting Friedman test and Nemeyi test. Then it is compared with the measures like precision, recall, accuracy and training time.
8) Later, the average rank and sum of squares will be calculated. By using the average rank, the Sum of Squared differences will be calculated.
9) After that, hypothesis test will be conducted to check whether all three algorithms are performing same or not. For alpha=0.05, n=10, k=3 from the given table using this parameter the critical value is 7.8. If that's the critical value then I reject hypothesis test, if that's not the critical value then we accept hypothesis. This mean all the three algorithms perform the same. And if the algorithms perform differently then we must conduct Nemenyi test. This test says which two algorithms perform differently.
10) Finally, this implementation results in comparison of the three different algorithms to know which algorithm will perform better.

III. Results

Results are obtained from the accuracy of State Vector Machine, logistic regression and KNN. These are shown in the following tables.

The null hypothesis gets rejected as all the above three algorithms perform equally. As Logistic regression and State Vector Machine are different in performance that exceeds critical difference and KNN and Logistic regression are different in performance that exceeds critical difference.

Accuracy:

Fold	LR	KNN	SVM
1.	0.9306 (1)	0.7419 (3)	0.7787 (2)
2.	0.9241 (1)	0.7744 (3)	0.7983 (2)
3.	0.9174 (1)	0.7804 (3)	0.8130 (2)
4.	0.9435 (1)	0.8217 (3)	0.8457 (2)
5.	0.9196 (1)	0.8087 (3)	0.8130 (2)
6.	0.9348 (1)	0.8239 (3)	0.8609 (2)
7.	0.9543 (1)	0.8196 (3)	0.8370 (2)
8.	0.9391 (1)	0.8283 (3)	0.8457 (2)
9.	0.8497 (1)	0.7233 (3)	0.7516 (2)
10.	0.8584 (1)	0.7625 (3)	0.8017 (2)
Average	0.9171 (1)	0.7884 (3)	0.8145 (2)
Average rank	1	3	2

Training time:

Fold	LR	KNN	SVM
1.	0.2303	0.0364	1.4274
2.	0.0628	0.0092	1.2884
3.	0.1019	0.0112	1.3725
4.	0.587	0.0158	1.3154
5.	0.1488	0.0091	1.3382
6.	0.0349	0.0100	1.3304
7.	0.0405	0.0103	1.2878
8.	0.0436	0.0080	1.3474
9.	0.0786	0.0160	1.3749
10.	0.0387	0.0161	1.3109
Average	0.9779	0.7413	2.1108
Average rank	2	1	3

F-measure:

Fold	LR	KNN	SVM
1.	0.9080	0.6571	0.7302
2.	0.9003	0.7095	0.7438
3.	0.8933	0.7089	0.7611
4.	0.9278	0.7574	0.8022
5.	0.8969	0.7582	0.7737
6.	0.9194	0.7793	0.8333
7.	0.9398	0.7462	0.7863
8.	0.9218	0.7799	0.8097
9.	0.8179	0.6718	0.7192
10.	0.8159	0.7030	0.7612
Average	0.8941	0.7271	0.7720
Average rank	1	3	2

Friedman's accuracy for accuracy is 20(> critical value 7.8)
From this, a different way of performance of algorithms is observed.
The average ranks between Logistic regression and K- nearest neighbors
is greater than the critical difference 1.047 (Namanya test). So,
we can see performance is not similar.

Friedman's statistic for F1-score is 20 (> critical value 7.8) From this
we can observe the different performance of algorithms. The average ranks
between logistic regression and K-nearest neighbors is greater than the critical
difference 1.047(Namanya test). So, we can see performance is not similar.

Friedman's statistic for training time is 30 (> critical value 7.8) From this
A different way of performance of algorithms is observed. The average ranks
between Logistic regression and State Vector Machine is greater than the[1]
critical difference 1.047(Namanya test). So, we can see performance is not
similar.

REFERENCES:

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