DV2599 - Assignment 2

1st Viktor Fransson
DVAMI22h
Blekinge Institute of Technology
Karlskrona, Sweden
vifr22@student.bth.se

2nd Tobias Gustafsson
DVAMI22h
Blekinge Institute of Technology
Karlskrona, Sweden
togu22@student.bth.se

I. Introduction

In this assignment, three different machine learning models were used for classifying e-mails as spam or not spam.

The three supervised classification algorithms chosen for the assignment were:

Random Forest - a model that trains an ensemble of tree models from bootstrap samples and random subspaces,

Logistic Regression - a model that combines a linear decision boundary with logistic calibration, trained discriminately by optimizing conditional likelihood- and,

Support Vector Classifier (SVC) - a kind of linear classifier that finds a decision boundary whose margin is as large as possible.

The description for each model was taken from *Machine learning: the art and science of algorithms that make sense of data* (Flach, 2012) [1] at pages 333, 296 and 22.

Before performing the ten-fold stratified cross-validation, the data was scaled using `StandardScaler` imported from the `sklearn` library.

II. Stratified Ten-fold Cross-validation

Each model was imported from the `sklearn` library. To evaluate the models, ten-fold stratified cross validation was implemented using `cross_validate` and `StratifiedKFold` from the `sklearn` library. The advantage of stratified cross validation is that it preserves similar class distribution for each fold [1].

For each model, a ten-fold stratified cross-validation was run. For each fold, fit time, accuracy score and f1 score were saved. The fit time was measured when accuracy was used as scoring measure. Then the average and standard deviation were calculated for each measurement and each model. The results for fit time can be seen in table I, accuracy in table II and f1 score in table III.

III. Friedman test

The Friedman test was used to evaluate the difference between the models. It works by ranking the performances for the models for each of the 10 folds (data sets). In this case the models gets a number 1 to 3 (can be same if the scores are equal), where 1 is the best score (lowest time, highest accuracy or highest f1-score) and 3 is the worst. For each model, an average score was also calculated. The results for the rankings are seen in table IV (fit time), V (accuracy score) and VI (f1-score).

TABLE I
Fit time for ten-fold stratified cross-validation (seconds)

Fold	*Random Forest*	*Logistic Regression*	*SVC*
1	0.5575	0.0277	0.1898
2	0.5722	0.0170	0.1445
3	0.5341	0.0170	0.1425
4	0.5481	0.0150	0.1583
5	0.5343	0.0156	0.1586
6	0.5366	0.0173	0.1510
7	0.5405	0.0150	0.1504
8	0.5422	0.0140	0.1601
9	0.5321	0.0143	0.1505
10	0.5391	0.0140	0.1415
avg	0.5437	0.0167	0.1547
stdev	0.0126	0.0041	0.0140

TABLE II
Accuracy score for ten-fold stratified cross-validation

Fold	*Random Forest*	*Logistic Regression*	*SVC*
1	0.9501	0.9306	0.9241
2	0.9478	0.9217	0.9413
3	0.9348	0.9196	0.9348
4	0.9478	0.9391	0.9370
5	0.9543	0.9239	0.9478
6	0.9565	0.9304	0.9522
7	0.9696	0.9522	0.9565
8	0.9761	0.9261	0.9413
9	0.8913	0.8543	0.8913
10	0.8630	0.8717	0.8696
avg	0.9391	0.9170	0.9296
stdev	0.0353	0.0302	0.0279

Then the Friedman statistic was calculated using the method presented in chapter 12.3 in (Flach, 2012) [1]. For each evaluation measure, three scores were calculated: the average rank (1), the sum of squared differences of ranks (2), and the sum of squared differences (3).

$$\overline{R} = \frac{1}{nk} \sum_{ij} R_{ij} \tag{1}$$

$$n \sum_{j} (R_j - \overline{R})^2 \tag{2}$$

TABLE III
F1 SCORE FOR TEN-FOLD STRATIFIED CROSS-VALIDATION

Fold	*Random Forest*	*Logistic Regression*	*SVC*
1	0.9352	0.9080	0.9025
2	0.9359	0.8977	0.9252
3	0.9138	0.8964	0.9157
4	0.9352	0.9222	0.9178
5	0.9441	0.9025	0.9326
6	0.9479	0.9144	0.9399
7	0.9607	0.9368	0.9419
8	0.9636	0.9045	0.9235
9	0.8616	0.8204	0.8611
10	0.8212	0.8300	0.8214
avg	0.9219	0.8933	0.9082
stdev	0.0457	0.0379	0.0383

TABLE IV
FRIEDMAN RANKS FOR FIT TIME

Data set	*Random Forest*	*Logistic Regression*	*SVC*
1	0.5575 (3)	0.0277 (1)	0.1898 (2)
2	0.5722 (3)	0.017 (1)	0.1445 (2)
3	0.5341 (3)	0.017 (1)	0.1425 (2)
4	0.5481 (3)	0.015 (1)	0.1583 (2)
5	0.5343 (3)	0.0156 (1)	0.1586 (2)
6	0.5366 (3)	0.0173 (1)	0.151 (2)
7	0.5405 (3)	0.015 (1)	0.1504 (2)
8	0.5422 (3)	0.014 (1)	0.1601 (2)
9	0.5321 (3)	0.0143 (1)	0.1505 (2)
10	0.5391 (3)	0.014 (1)	0.1415 (2)
avg rank	3.0	1.0	2.0

$$\frac{1}{n(k-1)}\sum_{ij}(R_{ij}-\overline{R})^2 \tag{3}$$

The Friedman statistic is calculated as the ratio between sum (2) and sum (3). The Friedman statistic is 20.0 for fit time, 8.3582 for accuracy score, and 8.6 for f1 score. To determine whether these algorithms perform significantly different, these values are compared to the critical value. The critical value for $\alpha = 0.05$, $N = 10$ and $k = 3$ is 6.200 [2]. The critical value is lower than all three calculated Friedman statistics, which indicates that for each one of the three measurements, at least one pair of algorithms do not perform equally.

IV. NEMENYI TEST

To determine which models perform differently in our three tests (fit time, accuracy score, f1 score), we use the Nemenyi test, as described in chapter 12.3 (Flach, 2012) [1]. In the Nemenyi test, for each measurement, the models are compared pairwise. The difference in average rank (*avg rank*) is calculated for each pair, and is compared to the *critical difference* (CD). If the difference between the two pairs exceeds CD, then the two models differ significantly. CD is calculated as (4):

$$CD = q_\alpha\sqrt{\frac{k(k+1)}{6n}} \tag{4}$$

TABLE V
FRIEDMAN RANKS FOR ACCURACY SCORE

Data set	*Random Forest*	*Logistic Regression*	*SVC*
1	0.9501 (1)	0.9306 (2)	0.9241 (3)
2	0.9478 (1)	0.9217 (3)	0.9413 (2)
3	0.9348 (2)	0.9196 (3)	0.9348 (2)
4	0.9478 (1)	0.9391 (2)	0.937 (3)
5	0.9543 (1)	0.9239 (3)	0.9478 (2)
6	0.9565 (1)	0.9304 (3)	0.9522 (2)
7	0.9696 (1)	0.9522 (3)	0.9565 (2)
8	0.9761 (1)	0.9261 (3)	0.9413 (2)
9	0.8913 (2)	0.8543 (3)	0.8913 (2)
10	0.863 (3)	0.8717 (1)	0.8696 (2)
avg rank	1.4	2.6	2.2

TABLE VI
FRIEDMAN RANKS FOR F1 SCORE

Data set	*Random Forest*	*Logistic Regression*	*SVC*
1	0.9352 (1)	0.908 (2)	0.9025 (3)
2	0.9359 (1)	0.8977 (3)	0.9252 (2)
3	0.9138 (2)	0.8964 (3)	0.9157 (1)
4	0.9352 (1)	0.9222 (2)	0.9178 (3)
5	0.9441 (1)	0.9025 (3)	0.9326 (2)
6	0.9479 (1)	0.9144 (3)	0.9399 (2)
7	0.9607 (1)	0.9368 (3)	0.9419 (2)
8	0.9636 (1)	0.9045 (3)	0.9235 (2)
9	0.8616 (1)	0.8204 (3)	0.8611 (2)
10	0.8212 (3)	0.83 (1)	0.8214 (2)
avg rank	1.3	2.6	2.1

Where $q_\alpha = 2.343$ for $k = 3$ and $\alpha = 0.05$ [1]. The critical distance is 1.0478.

For each measurement (fit time, accuracy score, f1 score) one pair of models exceeds the critical difference, meaning that the two models performed significantly different. For fit time, the pair Random forest/Logistic regression exceeds the critical distance. For accuracy score the pair Logistic regression/Random forest exceeds the critical distance. For F1 score the pair Logistic regression/Random forest exceeds the critical distance.

REFERENCES

[1] P. Flach, "Machine learning: the art and science of algorithms that make sense of data", Cambridge University Press, New York, 2012.

[2] Louise Martin, Raymond Leblanc and Nguyen Ky Toan, "Tables for the Friedman Rank Test", The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Vol. 21, No. 1 (Mar., 1993), pp. 39-43.