LAB 1 REPORT

A PREPRINT

Daniele cecca
Artificial Intelligence for Science and Technology
Milano Bicocca University
Supervised Learning

March 15, 2024

1 Introduction

The aim of this study is to conduct a comparison of various classifiers across multiple datasets. The goal is to evaluate the performance of different classifiers and identify any significant difference among them. We begin by introducing the classifiers used in the study. We present the datasets used by the classifiers. Then we discuss the cross validation method with a focus on our work. Subsequently, we present the results of the evaluations of the classifiers on the four datasets, followed by the computation and analysis of the rankings obtained by each classifier on each dataset. Additionally, we discuss the average rankings across all datasets and conduct statistical tests, including the Friedman test and the calculation of the Critical Difference (CD) value, to assess the significance of differences in classifier performance. Finally, we provide observations and insights based on the statistical analysis, highlighting the relative strengths and weaknesses of the classifiers.

2 Classifiers

The following classifiers were utilized for the comparison:

- **Support Vector Machine (SVM)** with a linear kernel (SVM Linear) Hyperparameters: Kernel Function - linear, Kernel Scale - 1
- **Support Vector Machine (SVM)** with a radial basis function (RBF) kernel (SVM RBF) Hyperparameters: Kernel Function - gaussian, Kernel Scale - 0.1
- **k-Nearest Neighbors (KNN)** Hyperparameters: Distance Metric - Euclidean, Number of Neighbors - 10
- **Decision Tree** Hyperparameters: Split Criterion - gdi, Maximum Number of Splits - 10

3 Data

In this study we used 4 datasets, each of them has 3 features: 2 predictors and 1 feature target.The features target assumes 2 values(1,2) which represents the two classes.

- **Dataset1** has 300 samples
- **Dataset2** has 300 samples
- **Dataset3** has 600 samples
- **Dataset4** has 600 samples

4 Cross-Validation

Cross-validation is a resampling technique used to assess the performance of a predictive model. It involves partitioning the dataset into multiple subsets, called folds, and iteratively training the model on a subset of the data while testing it

on the remaining data. This process is repeated multiple times, with each fold serving as both the training and testing set.

In our study, we employed a 5x2 cross-validation strategy, which involves splitting the dataset into 5 sets and repeating the process 2 times. This approach helps to ensure robustness in evaluating the classifiers' performance by reducing the variance associated with a single train-test split.

By using cross-validation, we obtain more reliable estimates of the classifiers' accuracies on unseen data, allowing for a more accurate comparison of their performance across different datasets.

5 Assesment of the models

The accuracies obtained by the classifiers on the four datasets are presented below:

Dataset	SVM Linear	SVM RBF	KNN	Decision Tree
1	1.0000	0.9987	1.0000	0.9967
2	0.8827	0.8740	0.8780	0.8773
3	0.6667	0.9373	0.9320	0.9310
4	0.5817	0.9767	0.9493	0.9733

The rankings obtained by the classifiers on each dataset are as follows:

Dataset	SVM Linear	SVM RBF	KNN	Decision Tree
1	1	3	2	4
2	1	4	2	3
3	4	1	2	3
4	4	1	3	2

The average rankings of the classifiers across all datasets are:

Table 1: Performance Rankings

Classifier	Rank
SVM Linear	2.50
SVM RBF	2.25
KNN	2.25
Decision Tree	3.00

N.B In this case, we used descending order in the ranking, and if two classifiers were in the same position, we just replicated the position. However, Friedman, to compute the ranks, used a more complex algorithm. Due to this, the means that Friedman have as outputs are slightly different.

6 Friedman Test

To determine if there is a statistically significant difference among the classifiers, we performed the Friedman test. The Friedman test is a non-parametric statistical test used to determine whether there are statistically significant differences in performance among multiple classifiers. It is based on rank sums and evaluates whether the mean ranks of the classifiers are equal across different datasets. A low p-value (< 0.05) indicates that at least one classifier significantly outperforms the others, leading us to reject the null hypothesis.

In general, it is computed by the following formula:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right]$$

In our study, we applied the Friedman test to assess the overall differences in performance among the classifiers across the datasets. Since we obtained a p-value near 1, we can conclude that the null hypothesis is validated. Thus, we can deduce that there will be no significant differences among the models.

The average rankings used by Friedman are:

Table 2: Performance Rankings

Classifier	Rank
SVM Linear	2.375
SVM RBF	2.75
KNN	2.875
Decision Tree	2.00

7 Critical Difference (CD)

Even though we have determined that there are no significant differences between the models, we also compute the Critical Difference (CD) value because it will be useful in the next tests. The CD value is used to identify significant pairwise differences between classifiers based on their average rankings. It is calculated using the Friedman test results and serves as a threshold for determining whether the differences in performance are statistically significant.

In general, it is computed by the following formula:

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

In our analysis, we computed the CD value using an alpha level of 0.05 (corresponding to a confidence level of 95

By observing the results, we can note that the p-value for each pairwise comparison is near 1, which means that no classifier outperforms another one.

Table 3: Comparisons between Models

Model 1	Model 2	p-value
1- SVM Linear	2- SVM rbf	0.9758
1- SVM Linear	3- KNN	0.9453
1- SVM Linear	4- Tree	0.9758
2- SVM rbf	3- KNN	0.9990
2- SVM rbf	4- Tree	0.8393
3- KNN	4- Tree	0.7661

We can also validate what we said above by the following graph, where we can see that all the mean value of rank are closed:

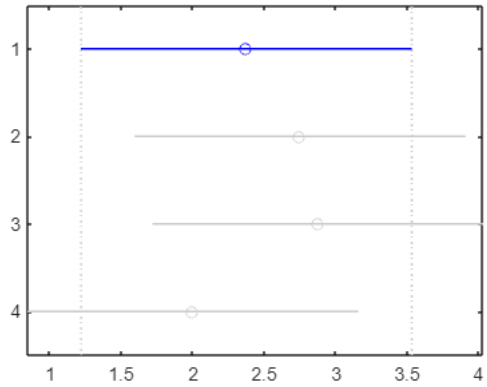

Figure 1: Critical Difference (CD) graph

8 Conclusion

In conclusion, from a statistical point of view we can say that there are not significant differences in performance among the classifiers.

But by only observing the small differences among the mean of ranks we can say that the KNN classifier demonstrated the best overall performance, followed closely by SVM RBF and then we have SVM Linear and Decision Tree.

References

[1] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. Disponibile su: `http://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf`. Accesso il 15 Marzo 2024.

[2] Stats. Multiple Comparison Test - MATLAB Multcompare. Disponibile su: `https://nl.mathworks.com/help/stats/multcompare.html`. Accesso il 15 Marzo 2024.

[3] X. Friedman's Test - MATLAB Friedman. Disponibile su: `https://nl.mathworks.com/help/stats/friedman.html#bubrp4t-stats`. Accesso il 15 Marzo 2024.